Physiol Genomics 7: 97–104, 2001.
First published October 2, 2001; 10.1152/physiolgenomics.00040.2001.

A compendium of gene expression in normal human tissues

LI-LI HSIAO,[1] FERNANDO DANGOND,[2] TAKUMI YOSHIDA,[1] ROBERT HONG,[1] RODERICK V. JENSEN,[3] JATIN MISRA,[4] WILLIAM DILLON,[1] KAILIN F. LEE,[1] KATHRYN E. CLARK,[1] PETER HAVERTY,[5] ZHIPING WENG,[6] GEORGE L. MUTTER,[7] MATTHEW P. FROSCH,[8] MARCY E. MACDONALD,[9] EDGAR L. MILFORD,[1] CHRISTOPHER P. CRUM,[10] RAPHAEL BUENO,[11] RICHARD E. PRATT,[12] MAMATHA MAHADEVAPPA,[13] JANET A. WARRINGTON,[13] GREGORY STEPHANOPOULOS,[4] GEORGE STEPHANOPOULOS,[4] AND STEVEN R. GULLANS[1]

[1]*Renal Division, Department of Medicine, and* [2]*Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston 02115;* [3]*Department of Physics, Wesleyan University, Middletown, Connecticut 06459;* [4]*Bioinformatics and Metabolic Engineering Laboratory, Department of Chemical Engineering, Massachusetts Institutes of Technology, Cambridge, Massachusetts 01890;* [5]*Bioinformatics Program and* [6]*Department of Biomedical Engineering, Bioinformatics Program, Boston University, Boston 02215;* [7]*Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston 02115;* [8]*Center for Neurologic Diseases, and Division of Neuropathology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston 02115;* [9]*Molecular Neurogenetics Unit, Massachusetts General Hospital, Charlestown Massachusetts 02129;* [10]*Division of Women's and Perinatal Pathology, Department of Pathology,* [11]*Division Thoracic Surgery, and* [12]*Cardiovascular Division, Brigham and Women's Hospital, Boston, Massachusetts 02115; and* [13]*Affymetrix, Inc., Santa Clara, California 95051*

Received 30 May 2001; accepted in final form 17 September 2001

Hsiao, Li-Li, Fernando Dangond, Takumi Yoshida, Robert Hong, Roderick V. Jensen, Jatin Misra, William Dillon, Kailin F. Lee, Kathryn E. Clark, Peter Haverty, Zhiping Weng, George L. Mutter, Matthew P. Frosch, Marcy E. MacDonald, Edgar L. Milford, Christopher P. Crum, Raphael Bueno, Richard E. Pratt, Mamatha Mahadevappa, Janet A. Warrington, Gregory Stephanopoulos, George Stephanopoulos, and Steven R. Gullans. A compendium of gene expression in normal human tissues. *Physiol Genomics* 7: 97–104, 2001. First published October 2, 2001; 10.1152/physiolgenomics.00040.2001.—This study creates a compendium of gene expression in normal human tissues suitable as a reference for defining basic organ systems biology. Using oligonucleotide microarrays, we analyze 59 samples representing 19 distinct tissue types. Of ~7,000 genes analyzed, 451 genes are expressed in all tissue types and designated as housekeeping genes. These genes display significant variation in expression levels among tissues and are sufficient for discerning tissue-specific expression signatures, indicative of fundamental differences in biochemical processes. In addition, subsets of tissue-selective genes are identified that define key biological processes characterizing each organ. This compendium highlights similarities and differences among organ systems and different individuals and also provides a publicly available resource (Human Gene Expression Index, the HuGE Index, **http://www.hugeindex.org**) for future studies of pathophysiology.

microarrays; human tissues; gene expression; bioinformatics

WITH THE RECENTLY ANNOUNCED completion of the human genome project (8, 37a), greater attention is now focused on defining the biological significance and functional properties of the ~30,000 human genes. Toward this end, a fundamental and primary objective is to define global patterns of gene expression that characterize human tissues in normal and disease states. DNA microarrays, along with other high-throughput approaches, can successfully elucidate expression patterns that distinguish disease states such as different types of cancers (2, 3, 6, 11, 15, 28). Individually, these distinguished genes are potential molecular markers or potential therapeutic targets for a disease process (5, 13, 14, 18, 23, 28, 38, 41). Establishment of baseline expression patterns in normal tissues is an essential element in accurate interpretation of those changes associated with pathological states.

In the present study we use oligonucleotide microarrays (GeneChip HuGeneFL) to analyze expression of 7,070 unique sequences in 59 tissue samples representing 19 healthy human tissue types. The purpose is to create a database that can serve as a reference or compendium of expression profiles for studies of human disease. Using a variety of statistical approaches, we identify gene expression patterns that characterize different tissue types. The results reveal striking quan-

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: S. R. Gullans, 65 Landsdowne, Cambridge, MA 02140 (E-mail: sgullans@rics.bwh.harvard.edu).

titative similarities and differences among tissues, even for those genes expressed constitutively.

METHODS AND MATERIALS

Tissue specimens. We obtained 59 human samples of 19 different tissue types, from 49 human individuals including 24 males and 25 females with median age of 63 and 50, respectively [for a table of demographic information, *Supplement 1*, please refer to the Supplementary Material[1] for this article, published online at the *Physiological Genomics* web site; this information is also available at the Human Gene Expression Index web site (the "HuGE Index") at **http://www.hugeindex.org**]. These were provided by tissue banks, surgical procedures or autopsies (Massachusetts General Hospital and Brigham and Women's Hospital) with appropriate Institutional Research Board consent. The specimens were immediately immersed in liquid nitrogen upon isolation. Each tissue was divided into matched fractions for RNA isolation and histological examination. Only those with normal histological examination were included in this study, and medical histories were not a criterion for exclusion.

Histology. Samples were fixed at room temperature in neutral pH phosphate-buffered 10% formalin, dehydrated in graded alcohols, and embedded in paraffin using an automated tissue processor. Four-millimeter-thick paraffin sections were rehydrated and stained routinely with hematoxylin and eosin. Light microscope examination was performed to confirm normal tissue morphology. Histological sections of the tissues are available at **http://www.hugeindex.org**.

RNA preparation for hybridization. Total RNA was isolated using Trizol solution (GIBC-BRL, Life Technologies, Rockville, MD). Seven micrograms of total RNA was used for amplification, and the amplified product was labeled with biotin following a procedure described previously (7, 25, 39). Briefly, double-stranded cDNA was synthesized using the SuperScript Choice System (GIBCO-BRL) and a T7-(dT)-24 primer (Geneset Oligos, La Jolla, CA). The cDNA was purified by phenol/chloroform/isoamyl alcohol extraction with Phase Lock Gel (5Prime → 3Prime, Boulder, CO) and concentrated by ethanol precipitation. In vitro transcription was performed to produce biotin-labeled cRNA using a BioArray HighYield RNA Transcript Labeling Kit (Affymetrix) according to the manufacturer's instructions. cRNA was linearly amplified with T7 polymerase. The biotinylated RNA was cleaned with RNeasy Mini kit (Qiagen, Valencia, CA).

Labeled cRNA, 20 μg, was fragmented and hybridized using the protocol described previously (25). Briefly, the hybridization mixture was incubated at 99°C for 5 min. followed by incubation at 45°C for 5 min. The hybridization was then carried out at 45°C for 16–18 h. After being washed, the array was stained with streptavidin-phycoerythrin (Molecular Probes, Eugene, OR), amplified by biotinylated anti-streptavidin (Vector Laboratories, Burlingame, CA), and then scanned on an HP Gene Array scanner. The intensity for each feature of the array was captured with Affymetrix GeneChip Software, according to standard Affymetrix procedures (25) by performing typical scaling (with target intensity of 100) and normalization for all probe sets.

Quality control of samples. Approximately 50% of total RNA collected from tissues were discarded secondary to unsatisfactory quality on a 1% agarose gel. Each probe array contains several prokaryotic genes (e.g., *bioB, bioC,* and *bioD* are genes of the biotin synthesis pathway from the bacteria *Escherichia coli*, *Cre* is the recombinase gene from P1 bacteriophage), which serve as hybridization controls. In addition, expression levels of 3′ to 5′ for both β-actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) were evaluated; the 3′/5′ ratio should be less than 3 according to the manufacturer's instructions. Data that failed to meet this criteria were excluded from analysis.

Statistical analysis. The Affymetrix GeneChip 3.1 Expression Analysis Algorithm present (P) or absent (A) calls were used to identify maintenance/housekeeping genes. All genes with a present call in at least one sample of each tissue type were included in the maintenance/housekeeping set [marginal (M) calls were conservatively treated as absent]. A hierarchical clustering algorithm (AGNES) (22) in the statistical analysis package SPLUS (37) was used to group the tissue samples using only the housekeeping genes. Using the "Manhattan" distance metric, variables standardized, and the "Ward" linkage algorithm, we found that the 451 housekeeping genes alone were sufficient to clearly group the different tissue types.

To identify tissue-selective genes, we used a two-tailed *t*-test to distinguish the gene expression levels in each tissue type from all other tissue samples at a 99.99% confidence level. The two-tailed *t*-test makes underlying assumptions about the distribution of the data, and this high confidence level was chosen to ensure that the list of tissue-selective genes obtained would still be reasonable, even though the assumptions may be met only in part (34). The tissue-selective genes obtained were ranked by their significance value, which determines the probability of observing a given level of discrimination for a gene by random chance. The lower the *P* value, the better the tissue-selective nature of the gene. A subset of 98 genes with the lowest *P* values, 14 from each tissue-selective subset from the brain, kidney, liver, lung, muscle, prostate, and vulva, were then used in a principal component analysis (PCA) to separate the tissue samples in PC space. Before performing the PCA, the data were autoscaled such that each gene had a mean of zero and unit standard deviation. This analysis was done using MATLAB. Finally, the coefficient of variation (CV = standard deviation/mean) for each tissue type was calculated to identify the tissue-variant genes.

RESULTS

Housekeeping or maintenance genes and their tissue-specific expression patterns. The Affymetrix GeneChip 3.1 Expression Analysis program uses a conservative call system to identify gene expression as "present" or "absent" among 7,070 unique sequences on each microarray. All 19 tissues in this study were analyzed, and a set of 451 genes with unique GenBank accession numbers (*Supplement 2*, published online at the *Physiological Genomics* web site and at **http://www.hugeindex.org**), 6.4% of 7,070 unique sequences, were identified as "present" in all 19 tissue types. This set of "housekeeping" or "maintenance" genes encodes proteins mediating a variety of basic cellular functions including intermediary metabolism, gene transcription, protein translation, cell signaling/communication, structure/motility, and other unclassified functions (1) (Fig. 1). Most of the ribosomal protein genes are included in this set. Overall, these housekeeping genes define basic cellular processes and could be used as a

[1]Supplementary Material (*supplements 1–5*) to this article is available online at **http://physiolgenomics.physiology.org/cgi/content/full/7/2/97/DC1.**

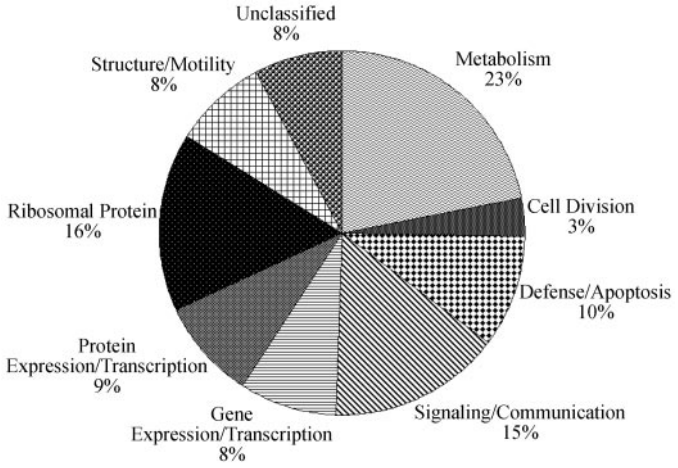

Fig. 1. The functional distribution of 451 maintenance genes in 59 samples representing 19 normal adult human tissue types.

reference standard when comparing gene expression studies.

A subset of 535 "housekeeping genes" was previously reported from 11 fetal and adult human tissue samples (39), and 358 genes are common to both lists. An important result in both studies is that the majority of the genes commonly considered to have a housekeeping function (e.g., β-actin and GAPDH) exhibit considerably variable expression levels from one tissue type to another. In fact, we found that quantitative expression profiles for the maintenance/housekeeping genes alone exhibit unique patterns for each specific tissue type. In particular, using a hierarchical clustering analysis of the 451 housekeeping genes, we successfully clustered different tissues according to tissue type (Fig. 2). We observed that all clusters were derived from two major branches: one branch contains the hematopoietic, reproductive, urinary tract, and gastrointestinal systems, while the other major branch contains the nervous, muscular, and liver systems.

We also identified the 15 most highly expressed genes among 451 (Table 1). They are primarily ribosomal proteins along with β-actin, GAPDH, and genes associated with defense/cell death (clusterin, metallothionein 2A). In addition, the CV for each of the maintenance/housekeeping genes was calculated, and the 15 genes with the highest and lowest CV were designated as most variable genes and most constant genes, respectively (Tables 2 and 3). Of note, both β-actin and

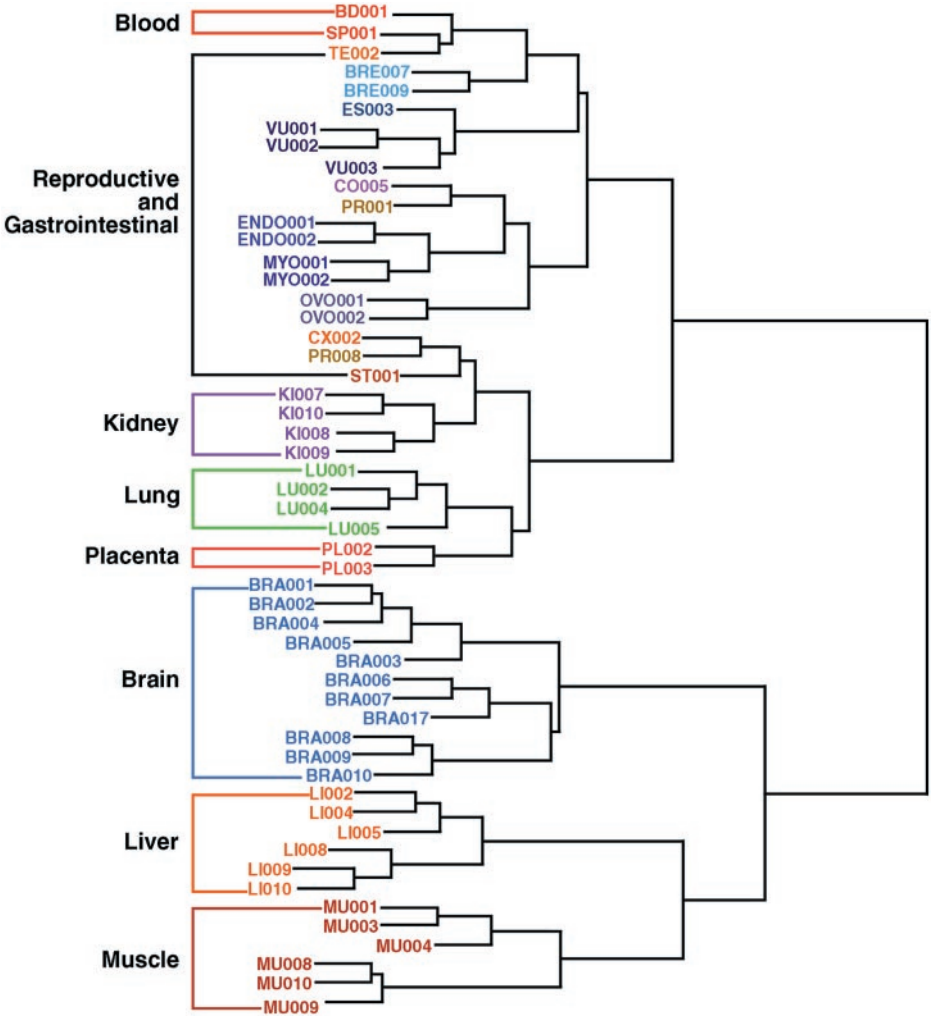

Fig. 2. Dendrogram revealing clustering profile of 19 tissue types using 451 maintenance genes. Cluster analysis performed using hierarchical Ward statistical method. BD, blood; BRA, brain; BRE, breast; CO, colon; CX, cervix; ENDO, endometrium; ES, esophagus; KI, kidney; LI, liver; LU, lung; MU, muscle; MYO, myometrium; OV, ovary; PL, placenta; PR, prostate; SP, spleen; ST, stomach; TE, testes; and VU, vulva. Each sample came from different individuals except for brain; BRA001–BRA007 were from different regions of same individual as well as LI005. Six samples, kidney (KI014, KI017), lung (LU014, LU018), and prostate (PR005, PR007), were omitted from hierarchical clustering analysis due to poor quality of chip data.

Table 1. *Fifteen most highly expressed housekeeping genes among 451*

Locus Link ID	Accession Number	Probe Set	Gene Name	Official Symbol	Mean Expression Level	SD	CV
6168	L06499	L06499_at	Ribosomal protein L37a	RPL37A	8638	4305	0.50
NA	Z70759	Z70759_at	Mitochondrial 16S rRNA	NA	7687	6245	0.81
6171	Z12962	Z12962_at	Ribosomal protein L41	RPL41	6915	3801	0.55
23117	D86974	D86974_at	KIAA0220	KIAA0220	6456	4377	0.68
1191	M63379	M63379_at	Clusterin	CLU	5497	6356	1.16
6175	M17885	M17885_at	Ribosomal protein, large, P0	RPLP0	5441	2738	0.50
60	M10277	M10277_s_at	Actin, beta	ACTB	5320	3054	0.57
NA	HG3214-HT3391	HG3214-HT3391_at	Metallopanstimulin 1	NA	5180	2897	0.56
6176	M17886	M17886_at	Ribosomal protein, large, P1	RPLP1	5047	2950	0.58
4502	V00594	V00594_s_at	Metallothionein 2A	MT2A	4989	6520	1.31
6218	M18000	M18000_at	Ribosomal protein S17	RPS17	4734	2022	0.43
7114	M17733	M17733_at	Thymosin, beta 4, X chromosome	TMSB4X	4637	2693	0.58
2597	X01677	X01677_f_at	Glyceraldehyde-3-phosphate dehydrogenase	GAPD	4614	2940	0.64
NA	HG2873-HT3017	HG2873-HT3017_at	Ribosomal protein L30 Homolog	NA	4592	1834	0.40
2512	M11147	M11147_at	Ferritin, light polypeptide	FTL	4557	3247	0.71

NA, not available; SD, standard deviation; and CV, coefficient of variation.

GAPDH, commonly assumed to have constant expression levels, were among the most variable genes. The 15 most constant genes among 451 maintenance/housekeeping genes could provide new standards for quantitative controls on all gene expression studies.

Identification of tissue-selective genes and class prediction. A two-tailed *t*-test was used to identify genes that are statistically highly expressed in a specific tissue type ($P < 0.0001$) using all 7,070 unique sequences from 59 samples. Our results reveal subsets of genes with unique GenBank accession numbers that are highly expressed only in brain (618 genes), kidney (91 genes), liver (277 genes), lung (75 genes), muscle (317 genes), prostate (46 genes), or vulva (101 genes). These are labeled "tissue-selective genes" (*Supplement 3*, published online at the *Physiological Genomics* web site and at **http://www.hugeindex.org**) as they are predominantly, not exclusively, expressed in one tissue type. Using 98 most selective genes (*Supplement 4*, published online at the *Physiological Genomics* web site and at **http://www.hugeindex.org**), we performed PCA to examine whether it was possible to discriminate among tissue types in three-dimensional expression space. As shown in Fig. 3, these "tissue-selective genes" can serve as templates for class prediction. For example, in brain the tissue-selective genes include those associated with myelin structure (e.g., myelin basic protein), with astrocytic differentia-

Table 2. *Fifteen most variable housekeeping genes among 451*

Locus Link ID	Accession Number	Probe Set	Gene Name	Official Symbol	Mean Expression Level	SD	CV
12	X68733	X68733_ma1_at	Serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	SERPINA3	664	1462	2.20
4256	X53331	X53331_at	Matrix Gla protein	MGP	1608	3254	2.02
2273	U60115	U60115_at	Four and a half LIM domains 1	FHL1	992	1963	1.98
1476	U46692	U46692_ma1_at	Cystatin B (stefin B)	CSTB	777	1313	1.69
5080	M93650	M93650_at	Paired box (PAX6) homolog	PAX6	324	545	1.68
2316	X53416	X53416_at	Filamin A, alpha (actin-binding protein-280)	FLNA	731	1160	1.59
3117	M34996	M34996_s_at	Major histocompatibility complex, class II, DQ alpha 1	HLA-DQA1	234	351	1.50
6281	M38591	M38591_at	S100 calcium-binding protein A10 (annexin II ligand, calpactin I, light polypeptide (p11))	S100A10	412	594	1.44
7832	U72649	U72649_at	BTG family, member 2 tyrosine 3 monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide	BTG2	593	790	1.33
7533	D78577	D78577_s_at		YWHAH	767	1005	1.31
4502	V00594	V00594_s_at	Metallothionein 2A	MT2A	4989	6520	1.31
6324	L10338	L10338_s_at	Sodium channel, voltage-gated, type I, beta polypeptide	SCN1B	552	710	1.29
3983	D31883	D31883_at	Actin binding LIM protein 1	ABLIM	1006	1272	1.26
8337	L19779	L19779_at	H2A histone family, member O	H2AFO	289	365	1.26
8926	J04615	J04615_at	SNRPN upstream reading frame	SNURF	1267	1533	1.21

Table 3. *Fifteen most constant housekeeping genes among 451*

Locus Link ID	Accession Number	Probe Set	Gene Name	Official Symbol	Mean Expression Level	SD	CV
5708	D78151	D78151_at	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 2	PSMD2	576	163	0.28
5690	D26599	D26599_at	Proteasome (prosome, macropain) subunit, beta type, 2	PSMB2	587	169	0.29
2873	U20285	U20285_at	G protein pathway suppressor 1	GPS1	349	105	0.30
5691	D26598	D26598_at	Proteasome (prosome, macropain) subunit, beta type, 3	PSMB3	476	144	0.30
821	L10284	L10284_at	Calnexin	CANX	709	223	0.32
375	M84332	M84332_at	ADP-ribosylation factor 1	ARF1	873	282	0.32
6944	D43642	D43642_at	Transcription factor-like 1	TCFL1	484	157	0.33
3020	M11353	M11353_at	H3 histone, family 3A	H3F3A	1857	630	0.34
9791	D14694	D14694_at	Phosphatidylserine synthase 1	PTDSS1	343	119	0.35
3183	M16342	M16342_at	Heterogeneous nuclear ribonucleoprotein C (C1/C2)	HNRPC	188	65	0.35
3735	D31890	D31890_at	Lysyl-tRNA synthetase	KARS	398	140	0.35
5879	D25274	D25274_at	Ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)	RAC1	713	257	0.36
14	M95627	M95627_at	Angio-associated, migratory cell protein	AAMP	365	132	0.36
2098	M13450	M13450_at	Esterase D/formylglutathione hydrolase	ESD	363	131	0.36
6917	M81601	M81601_at	Transcription elongation factor A (SII), 1	TCEA1	241	87	0.36

tion (e.g., glial fibrillary acidic protein), with synaptic reorganization (e.g., calcium channel, voltage-dependent β2, calmodulin 3, and GAP-43), and with neurotransmission (e.g., GABA receptor and glial high-affinity glutamate transporter). Among the tissue-selective genes (*Supplement 3*) we also identify smaller subsets of "tissue-specific" genes, defined statistically as $P = 0$ and having no overlap in expression level with any other tissue.

Kidney-selective genes include those known to be highly expressed in this organ such as uromodulin (Tamm-Horsfall glycoprotein), α-enolase, and ion transporters (e.g., β1-subunit of Na^+-K^+-ATPase, Na-Cl electroneutral thiazide-sensitive cotransporter, K-inwardly-rectifying channel, bumetanide-sensitive Na-K-2Cl cotransporter, amiloride-sensitive epithelial sodium channel, and amiloride binding protein 1). In addition, hydroxysteroid (11-β) dehydrogenase 2 (11β-HSD2), a gene that inactivates glucocorticoids and prevents them from binding to the nonselective mineralocorticoid receptor, is also highly expressed. In the kidney, it is this NAD-dependent high-affinity isoform which is thought to endow specificity on the receptor comprising nature of an autocrine protector of the mineralocorticoid receptor and play an important role in cardiovascular homeostatic mechanism (24).

As anticipated, the liver-selective genes include those associated with the coagulation pathway (e.g., factors II, V, VII, IX–XII, fibrinogen, plasminogen, protein S, and antithrombin III), complement pathway (e.g., C2, C4, C5, C8, C9), alcohol metabolism (e.g., alcohol dehydrogenase), lipid process (e.g., apolipoproteins), bile metabolism (e.g., bile acid CoA:amino acid *N*-acyltransferase), antitrypsin member 8, and xenobiotic metabolism (e.g., cytochrome *P*-450). Additionally, serum amyloid A1 and A4, constitutive for amyloid fibril formation, angiogenin, ribonuclease, RNase for angiogenesis, α1-glycine amidinotransferase for creatine biosynthesis, cysteine dioxygenase, type I for cysteine metabolism, and genes associated with growth, such as insulin-like growth factor, growth hormone receptor, hepatocyte growth factor (HGF) activator, are highly expressed in liver as well. Unlike 11β-HSD2, a

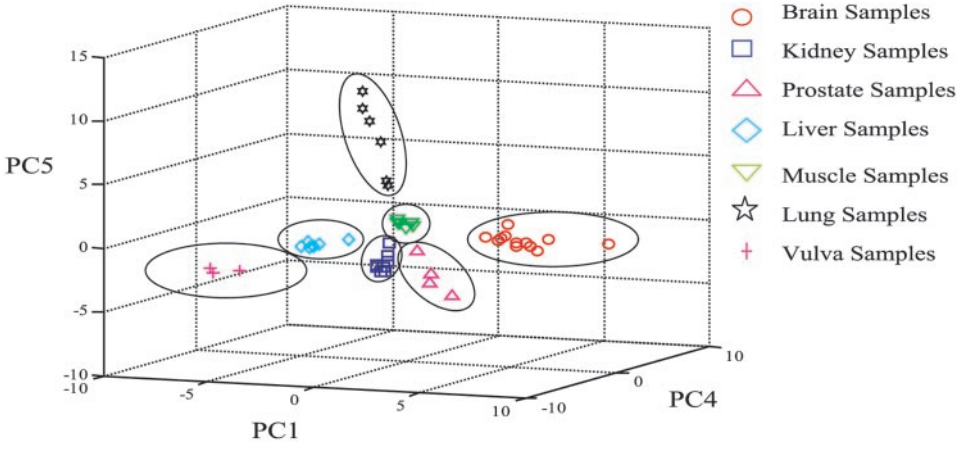

Fig. 3. Separation of the 7 tissue groups in 3-dimensional principal component (PC) space by applying principal component analysis (PCA) on 98 tissue-specific genes. Fourteen tissue-selective genes for each class of tissues, such as the brain, kidney, and other classes of samples were chosen to enhance the separation in PC space. When these 98 genes are used, the discrimination of the tissues can be observed in the 1st, 4th, and 5th PCs. Together, these 3 PCs capture ~40% of the information in the data.

gene specific in kidney, 11β-HSD1 is the gene in liver involved in steroid metabolism.

The lung-selective genes include those associated with extracellular matrix (e.g., pulmonary surfactant associated protein), HLA/cytokine (e.g., MHC II, γ-interferon inducible protein 30) and others (von Willebrand factor, claudin 5, palmitoyl-protein thioesterase 2, mannose receptor, lung cytochrome *P*-450). The muscle-selective genes include those associated with the cytoskeleton (e.g., actin, α1, actinin α2–3), contraction (e.g., tropomyosin, troponin, myosin), mitochondria (e.g., cytochrome C-1, ubiquitin, creatine kinase), and metabolism of glucose, glycogen, and lipids (e.g., lactate dehydrogenase, phosphoglucomutase 1, carnitine palmitoyltransferase). Furthermore, carbonic anhydrase III for CO_2 metabolism, creatine kinase, mitochondrial 2 (sarcomeric) for energy transduction, and gene for thermal regulation (neurotrophic tyrosine kinase, receptor, type 1) are also highly expressed in muscle.

The prostate-selective genes include those that are associated with hormones (e.g., prostate secretory protein), redox pathways (e.g., prostatic acid phosphatase, aldehyde dehydrogenase 6), cytoskeleton (actin-binding protein-278), and others (e.g., prostate-specific antigen, T-cell receptor-γ, TGF-β3, estrogen regulated LIV-1 protein). The vulva-selective genes include those associated with the cytoskeleton (e.g., keratin, ladinin, loricrin), extracellular matrix (e.g., desmoplakin 1, profilaggrin, epican, connexin 26, galectin 7, desmocollin), and hair follicle-related protein (e.g., basic/acidic hair keratin).

Identification of variant genes within a tissue. Another question of significant interest is whether there are genes whose tissue-specific expression is highly variable between different individuals. This was done by calculating the CV for genes called "present" in all samples. Figure 4 shows a histogram depicting the distribution of CV among different kidney specimens. The mean CV for the distribution was 0.31 with a standard deviation of 0.25. The genes with CV score greater than two standard deviations away from the mean are highlighted, indicating those that are most variable in kidney. These transcripts include several known to be associated with disease phenotypes. For example, the Na-Cl electroneutral thiazide-sensitive cotransporter is the target of a major antihypertensive diuretic, and mutations in this gene can cause Gitelman's syndrome, an autosomal recessive disease characterized by diverse abnormalities in electrolyte homeostasis (7, 36). In addition, aldose reductase plays a key role in the diabetic complications of kidney, nerve, and retina (10, 19, 29–31). We also observed similar distribution patterns of CV for brain, liver, lung, muscle, and vulva and identified small subsets (<2%) of genes that are highly variable (*Supplement 5*, published online at the *Physiological Genomics* web site and at **http://www.hugeindex.org**). In lung, the most variant genes include integrin-β2, which has been shown to predispose individuals to recurrent bacterial infections (16, 20), and antileukoproteinase, which is involved in several chronic and acute diseases of the respiratory tract (4, 35). In liver, the most variant genes include insulin-like growth factor 2, a putative susceptibility factor for obesity (9, 12, 32); fibrinogen-γ, defects of which are a cause of thrombophilia (26, 27); and hepatic lipase, a complete deficiency of which causes coronary atherosclerosis and premature dyslipidemia (17, 33). Each of the other 13 tissue types contains samples from less than 3 different individuals. Therefore, CVs were not calculated.

DISCUSSION

In 1965, Watson et al. (40) defined the housekeeping genes as those genes that are "always expressed" in

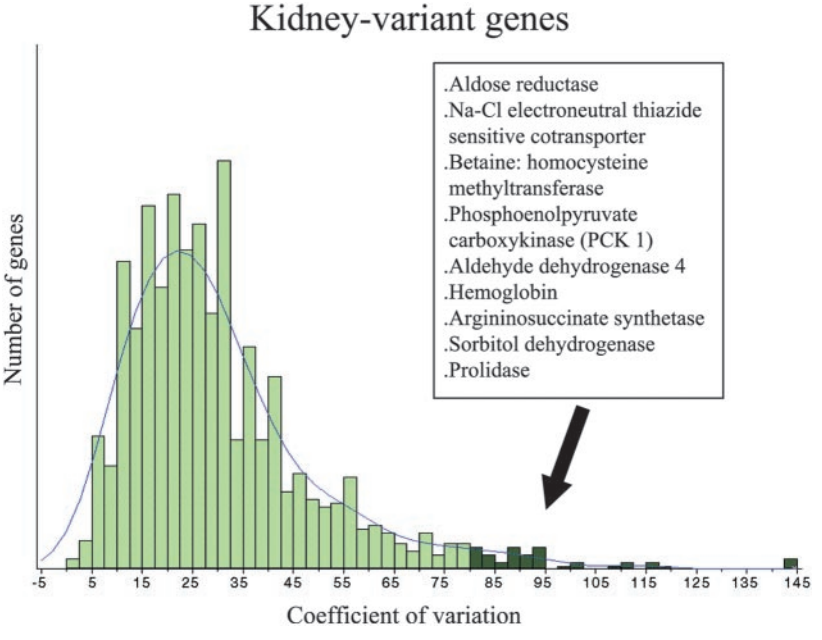

Fig. 4. Histogram showing the distribution of genes present in all 4 normal human kidney tissues. Coefficient of variation (CV) scores were calculated for each of the present genes. The *x*-axis represents the CV score, and the *y*-axis represents the number of genes. The mean CV for the distribution was 0.31 with a standard deviation of 0.25. The genes with CV score greater than 2 standard deviations away from the mean are highlighted.

every tissue to maintain cellular functions. This study is the largest quantitative survey, evaluating ~7,000 expressed sequences from 19 normal adult human tissue types. We identified a subset of 451 genes expressed in all normal adult human tissue types. This result supplements a previous report of 535 genes that were expressed in 11 fetal and adult tissues (39). Also, 358 of these maintenance/housekeeping genes are common to both lists. Functional annotation revealed that these genes participate in many active cellular processes.

We also found that expression of many of these maintenance/housekeeping genes is highly variable. In particular, we report here that these maintenance/housekeeping genes alone contain "tissue-specific" expression patterns (Fig. 2), which may be used to distinguish an individual tissue type. These results suggest that the gene expression patterns of maintenance/housekeeping genes reflect intrinsic differences among the individual tissues, most likely related to differences in metabolic activity and cytoarchitecture. The ability of housekeeping genes to define different biological states suggests that they may be suitable for distinguishing different disease states as well. In a practical sense, these genes could be used as standard controls on all gene expression studies to facilitate data comparison among laboratories and across platforms.

We identified subsets of genes that are highly expressed in one tissue type but not in others. We labeled these "tissue-selective genes" rather than "tissue-specific genes," since very few were expressed in only a single tissue type and a number of human tissues were not included in our analysis. These tissue-selective genes, ranging from 75 to 621 genes per tissue, have enough power to provide class prediction using a three-dimensional PCA. Additionally, these subsets of genes are found to be closely associated with the major functions carried out by each specific tissue type, e.g., genes related to myelin proteins and glial differentiation in brain; genes involved in coagulation and complement pathways in liver; genes associated with channels and transporters in kidney; and genes for pulmonary surfactant proteins in lung. We suggest that these genes may represent potential "signature" genes for the specific tissues with the important caveat that more tissues need to be sampled (e.g., endocrine system) to refine the "tissue-selective" fingerprints. Furthermore, ongoing efforts to deduce the functions of orphan genes will benefit from defining those that have tissue-selective expression patterns, as this will highlight a limited number of biological processes that should be considered.

Although all samples used in this study are normal tissues from a histological perspective, one important observation from our study is the demonstration that, for a given tissue, different individuals have a small set of genes with highly variable expression. Logically, tissues from either biopsy or autopsy often contain multiple cell types, which are in various states. It is conceivable that the variations are due to differences in the cell types and their states when the tissues were collected. In addition, the differences of age, gender, underlying health, and medications may also play roles in the variation. Further studies will be needed to address these issues. Even so, the presence of tissue-variant genes is consistent with the notion that the genotypes and the inherent plasticity of human tissues of an individual may contribute to gene-specific expression.

We thank Yangxi Wang, Mani Khounviengsay, Nathan Best, and Ken Auerbach for web site design. We also thank Frank Haluska and Mohammed Miri for assistance with tissue acquisition. We acknowledge Lynn Mills, a high school biology teacher, for inspiring the name of the HuGE Index.

This work was supported by the Merck Genome Research Institute. In addition, support was provided by National Institutes of Health Grants DK-36031 and DK-58849 (to S. R. Gullans), DK-09987 (to L.-L. Hsiao), CA-80084 (to F. Dangond), NS-16367 (to M. Mahadevappa), and DK-58533 (to Gregory Stephanopoulos). This work was also partially supported by Grants DE-FG02-94ER-14487 and DE-FG02-99ER-15015 (to Gregory Stephanopoulos) from the Engineering Research Program of the Office of Basic Energy Science at the Dept. of Energy, by the BSG foundation (to R. Bueno), and by Integrative Graduate Education and Research Traineeship 9870710 (to P. Haverty).

REFERENCES

1. **Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, and et al.** Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377: 3–174, 1995.
2. **Alaiya AA, Franzen B, Hagman A, Silfversward C, Moberger B, Linder S, and Auer G.** Classification of human ovarian tumors using multivariate data analysis of polypeptide expression patterns. *Int J Cancer* 86: 731–736, 2000.
3. **Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PD, and Staudt LM.** Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503–511, 2000.
4. **Ameshima S, Ishizaki T, Demura Y, Imamura Y, Miyamori I, and Mitsuhashi H.** Increased secretory leukoprotease inhibitor in patients with nonsmall cell lung carcinoma. *Cancer* 89: 1448–1456, 2000.
5. **Bubendorf L, Kolmer M, Kononen J, Koivisto P, Mousses S, Chen Y, Mahlamaki E, Schraml P, Moch H, Willi N, Elkahloun AG, Pretlow TG, Gasser TC, Mihatsch MJ, Sauter G, and Kallioniemi OP.** Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. *J Natl Cancer Inst* 91: 1758–1764, 1999.
6. **Camp RL, Charette LA, and Rimm DL.** Validation of tissue microarray technology in breast carcinoma. *Lab Invest* 80: 1943–1949, 2000.

6a. **Celera Genomics.** The sequence of the human genome. *Science* 291: 1304–1351, 2001.

7. **Colussi G, Rombola G, Brunati C, and De Ferrari ME.** Abnormal reabsorption of Na^+/Cl^- by the thiazide-inhibitable transporter of the distal convoluted tubule in Gitelman's syndrome. *Am J Nephrol* 17: 103–111, 1997.
8. **Consortium of the IHGS.** Initial sequencing and analysis of the human genome. *Nature* 409: 813–958, 2001.
9. **Devedjian JC, George M, Casellas A, Pujol A, Visa J, Pelegrin M, Gros L, and Bosch F.** Transgenic mice overexpressing insulin-like growth factor-II in beta cells develop type 2 diabetes. *J Clin Invest* 105: 731–740, 2000.
10. **Dunlop M.** Aldose reductase and the role of the polyol pathway in diabetic nephropathy. *Kidney Int* 58, *Suppl* 77: S3–S12, 2000.

11. **Elek J, Park KH, and Narayanan R.** Microarray-based expression profiling in prostate tumors. *In Vivo* 14: 173–182, 2000.
12. **Frystyk J, Skjaerbaek C, Vestbo E, Fisker S, and Orskov H.** Circulating levels of free insulin-like growth factors in obese subjects: the impact of type 2 diabetes. *Diabetes Metab Res Rev* 15: 314–322, 1999.
13. **Glynne R, Akkaraju S, Healy JI, Rayner J, Goodnow CC, and Mack DH.** How self-tolerance and the immunosuppressive drug FK506 prevent B-cell mitogenesis. *Nature* 403: 672–676, 2000.
14. **Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, and Lander ES.** Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537, 1999.
15. **Gray JW and Collins C.** Genome changes and gene expression in human solid tumors. *Carcinogenesis* 21: 443–452, 2000.
16. **Harris ES, Shigeoka AO, Li W, Adams RH, Prescott SM, McIntyre TM, Zimmerman GA, and Lorant DE.** A novel syndrome of variant leukocyte adhesion deficiency involving defects in adhesion mediated by beta(1) and beta(2) integrins. *Blood* 97: 767–776, 2001.
17. **Heller F, Descamps O, Hondekijn JC, and Desager JP.** Atorvastatin and low-density lipoprotein subfractions profile in mixed hyperlipidaemia: a contributory effect of reduced hepatic lipase activity? *Ann Clin Biochem* 36: 788–789, 1999.
18. **Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, and Davis RW.** Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 94: 2150–2155, 1997.
19. **Hilz MJ, Marthol H, and Neundorfer B.** Diabetic somatic polyneuropathy. Pathogenesis, clinical manifestations and therapeutic concepts. *Fortschr Neurol Psychiatr* 68: 278–288, 2000.
20. **Hogg N, Stewart MP, Scarth SL, Newton R, Shaw JM, Law SK, and Klein N.** A novel leukocyte adhesion deficiency caused by expressed but nonfunctional β2 integrins Mac-1 and LFA-1. *J Clin Invest* 103: 97–106, 1999.
22. **Kauffman L and Rouseeuw PJ.** *Finding Groups in Data*. New York: Wiley, 1990.
23. **Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, and Meltzer PS.** Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 58: 5009–5013, 1998.
24. **Krozowski Z, MaGuire JA, Stein-Oakley AN, Dowling J, Smith RE, and Andrews RK.** Immunohistochemical localization of the 11β-hydroxysteroid dehydrogenase type II enzyme in human kidney and placenta. *J Clin Endocrinol Metab* 80: 2203–2209, 1995.
25. **Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, and Brown EL.** Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14: 1675–1680, 1996.
26. **Marchi R, Lundberg U, Grimbergen J, Koopman J, Torres A, de Bosch NB, Haverkate F, and Arocha Pinango CL.** Fibrinogen Caracas V, an abnormal fibrinogen with an Aalpha 532-Ser→Cys substitution associated with thrombosis. *Thromb Haemost* 84: 263–270, 2000.
27. **Marchi R, Mirshahi SS, Soria C, Mirshahi M, Zohar M, Collet JP, de Bosch NB, Arocha-Pinango CL, and Soria J.** Thrombotic dysfibrinogenemia. Fibrinogen "Caracas V" relation between very tight fibrin network and defective clot degradability. *Thromb Res* 99: 187–193, 2000.
28. **Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP, and Sauter G.** Identification of prognostic parameters for renal cell carcinoma by cDNA arrays and cell chips. *Verh Dtsch Ges Pathol* 83: 225–232, 1999.
29. **Nishikawa T, Edelstein D, and Brownlee M.** The missing link: a single unifying mechanism for diabetic complications. *Kidney Int* 58, *Suppl* 77: S26–S30, 2000.
30. **Nishimura C, Matsuura Y, Kokai Y, Akera T, Carper D, Morjana N, Lyons C, and Flynn TG.** Cloning and expression of human aldose reductase. *J Biol Chem* 265: 9788–9792, 1990.
31. **Obrosova IG and Fathallah L.** Evaluation of an aldose reductase inhibitor on lens metabolism, ATPases and antioxidative defense in streptozotocin-diabetic rats: an intervention study. *Diabetologia* 43: 1048–1055, 2000.
32. **O'Dell SD, Bujac SR, Miller GJ, and Day IN.** Associations of IGF2 ApaI RFLP and INS VNTR class I allele size with obesity. *Eur J Hum Genet* 7: 821–827, 1999.
33. **Pihlajamaki J, Karjalainen L, Karhapaa P, Vauhkonen I, Taskinen MR, Deeb SS, and Laakso M.** G-250A substitution in promoter of hepatic lipase gene is associated with dyslipidemia and insulin resistance in healthy control subjects and in members of families with familial combined hyperlipidemia. *Arterioscler Thromb Vasc Biol* 20: 1789–1795, 2000.
34. **Rice JA.** *Mathematical Statistics and Data Analysis* (2nd ed.). Belmont, CA: Duxbury Press, 1995.
35. **Sallenave JM, Donnelly SC, Grant IS, Robertson C, Gauldie J, and Haslett C.** Secretory leukocyte proteinase inhibitor is preferentially increased in patients with acute respiratory distress syndrome. *Eur Respir J* 13: 1029–1036, 1999.
36. **Simon DB, Nelson-Williams C, Bia MJ, Ellison D, Karet FE, Molina AM, Vaara I, Iwata F, Cushner HM, Koolen M, Gainza FJ, Gitleman HJ, and Lifton RP.** Gitelman's variant of Bartter's syndrome, inherited hypokalaemic alkalosis, is caused by mutations in the thiazide-sensitive Na-Cl cotransporter. *Nat Genet* 12: 24–30, 1996.
37. **Venables WN and Ripley BD.** *Modern Applied Statistics with S-Plus* (2nd ed.). New York: Springer, 1997.
38. **Wang K, Gan L, Jeffery E, Gayle M, Gown AM, Skelly M, Nelson PS, Ng WV, Schummer M, Hood L, and Mulligan J.** Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* 229: 101–108, 1999.
39. **Warrington JA, Nair A, Mahadevappa M, and Tsyganskaya M.** Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* 2: 143–147, 2000.
40. **Watson JD, Hopkins NH, Roberts JW, Steitz JA, and Weiner AM.** *Molecular Biology of the Gene* (4th ed.). Benjamin/Cummings, 1987, vol. 1, p. 704.
41. **Whitney LW, Becker KG, Tresser NJ, Caballero-Ramos CI, Munson PJ, Prabhu VV, Trent JM, McFarland HF, and Biddison WE.** Analysis of gene expression in multiple sclerosis lesions using cDNA microarrays. *Ann Neurol* 46: 425–428, 1999.