Resource

An atlas of human gene expression from massively parallel signature sequencing (MPSS)

C. Victor Jongeneel,[1,6] Mauro Delorenzi,[2] Christian Iseli,[1] Daixing Zhou,[4] Christian D. Haudenschild,[4] Irina Khrebtukova,[4] Dmitry Kuznetsov,[1] Brian J. Stevenson,[1] Robert L. Strausberg,[5] Andrew J.G. Simpson,[3] and Thomas J. Vasicek[4]

[1]Office of Information Technology, Ludwig Institute for Cancer Research, and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [2]National Center for Competence in Research in Molecular Oncology, Swiss Institute for Experimental Cancer Research (ISREC) and Swiss Institute of Bioinformatics, 1066 Epalinges, Switzerland; [3]Ludwig Institute for Cancer Research, New York, New York 10012, USA; [4]Solexa, Inc., Hayward, California 94545, USA; [5]The J. Craig Venter Institute, Rockville, Maryland 20850, USA

We have used massively parallel signature sequencing (MPSS) to sample the transcriptomes of 32 normal human tissues to an unprecedented depth, thus documenting the patterns of expression of almost 20,000 genes with high sensitivity and specificity. The data confirm the widely held belief that differences in gene expression between cell and tissue types are largely determined by transcripts derived from a limited number of tissue-specific genes, rather than by combinations of more promiscuously expressed genes. Expression of a little more than half of all known human genes seems to account for both the common requirements and the specific functions of the tissues sampled. A classification of tissues based on patterns of gene expression largely reproduces classifications based on anatomical and biochemical properties. The unbiased sampling of the human transcriptome achieved by MPSS supports the idea that most human genes have been mapped, if not functionally characterized. This data set should prove useful for the identification of tissue-specific genes, for the study of global changes induced by pathological conditions, and for the definition of a minimal set of genes necessary for basic cell maintenance. The data are available on the Web at http://mpss.licr.org and http://sgb.lynxgen.com.

[Supplemental material is available online at www.genome.org. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: A. Delaney.]

As a rule, adult human organs and tissues perform highly specialized tasks, and contain cell types that have gone through an extensive differentiation program. Cells belonging to different tissues can be distinguished morphologically, functionally, and biochemically. Differentiation is driven largely by changes in the transcriptional program of the cells, through regulatory and epigenetic events. Therefore, the availability of comprehensive snapshots of the transcriptomes of cell populations from fully differentiated tissues should give us valuable information about the genes whose expression is necessary to maintain their specialized functions, as well as those that are necessary to all living cells. We have shown previously that massively parallel signature sequencing (MPSS) is a technique that can provide such a picture, at least for the vast majority of human transcripts (Jongeneel et al. 2003). MPSS is unlike microarrays, where issues of array design, cross-hybridization and reproducibility limit the coverage and dynamic range of the assay. MPSS also has the advantage that it samples the transcripts present in an mRNA population in an essentially unbiased fashion.

We have analyzed pooled RNA samples isolated from 32 human tissues, and were able to document the patterns of expression of 18,667 genes. The identities and relative expression levels of these genes give valuable insights into the specialized functions performed by fully differentiated tissues, and into the gene products required to maintain them. Moreover, these data largely define the complement of genes expressed in a variety of normal tissues, and thus a backdrop against which pathological changes can be detected and analyzed.

[6]Corresponding author.
E-mail Victor.Jongeneel@licr.org; fax 41-21-6924065.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.4041005.

Results

Depth of coverage and mapping of signatures

mRNA populations extracted from 23 different non-CNS organs and from nine different CNS areas (Table 1) were subjected to MPSS analysis (Brenner et al. 2000a), with 1.3×10^6 to 6×10^6 signatures being generated in two reading phases for each sample. The cDNA libraries attached to microbeads were produced using the original Megaclone protocol (Brenner et al. 2000b), which includes the amplification by PCR of the entire region between the poly(A) tail and the first DpnII site on the cDNA. Six batches of loaded beads were used for sequencing each sample, each representing an aliquot of 1.6×10^5 molecules drawn from an initial library with a complexity of 4×10^7 to 4×10^8 independent cDNA/vector ligations; therefore, the maximum complexity of the sampled population is 9.6×10^5. Only signatures seen in at least two independent sequencing runs and present at a minimum number of three transcripts per

Table 1. Tissues sampled and annotation process

Tissue	Clones	Signatures	Mapped	% mapped	Genes	In loci	Reverse	Genome	Unmapped
Adrenal gland	3.4	28,610	17,352	60.7%	9761	7166	3403	434	255
Bladder	2.5	25,795	14,092	54.6%	8284	6787	4320	480	116
Bone marrow	2.5	21,392	11,352	53.1%	7182	5492	3592	264	692
Brain (amygdala)	3.1	30,777	17,886	58.1%	10,168	8370	3594	693	234
Brain (caudate nucleus)	3.0	28,749	16,940	58.9%	9948	7509	3446	646	208
Brain (cerebellum)	1.8	27,372	13,413	49.0%	8183	9152	3880	717	210
Brain (corpus callosum)	3.5	31,526	18,302	58.1%	10,210	8679	3632	656	257
Brain (fetal)	4.7	27,030	13,699	50.7%	8447	9110	3158	765	298
Brain (hypothalamus)	2.2	27,474	16,710	60.8%	9867	6371	3384	585	424
Brain (thalamus)	1.5	25,936	15,680	60.5%	9389	6719	2800	551	186
Heart	3.6	27,335	16,906	61.8%	9379	5689	4147	369	224
Kidney	2.5	20,996	12,366	58.9%	7631	4425	3647	368	190
Lung	3.2	29,071	17,344	59.7%	9836	7384	3713	381	249
Mammary gland	2.6	20,014	13,162	65.8%	8351	4378	1979	321	174
Pancreas	2.4	11,634	8280	71.2%	5845	1958	1192	113	91
Pituitary gland	2.2	25,997	15,381	59.2%	9220	6255	3568	591	202
Placenta	5.3	16,344	10,236	62.6%	6631	3790	1856	283	179
Prostate	3.2	21,097	12,521	59.3%	7941	5427	2584	338	227
Retina	4.3	27,152	16,504	60.8%	9891	6368	3380	537	363
Salivary gland	3.6	17,598	11,706	66.5%	7540	4065	1476	223	128
Small intestine	5.0	29,599	18,084	61.1%	10,286	7902	2898	485	230
Spinal cord	4.2	30,367	18,920	62.3%	10,632	7214	3437	585	211
Spleen	3.1	34,956	19,988	57.2%	10,594	9877	4198	579	314
Stomach	3.1	13,992	9544	68.2%	6537	2995	1205	161	87
Testis	3.4	39,127	22,731	58.1%	12,267	9549	4944	1490	413
Thymus	5.5	36,250	19,871	54.8%	10,470	11,215	4001	742	421
Thyroid	6.0	29,210	17,596	60.2%	9870	7190	3556	487	381
Trachea	3.0	27,516	17,823	64.8%	10,201	6620	2333	440	300
Uterus	4.9	34,187	20,085	58.8%	10,737	9528	3753	529	292
Colon	1.3	13,982	9293	66.5%	6149	3267	1090	238	94
Monocytes	1.9	22,351	13,674	61.2%	7533	5125	2764	436	352
Peripheral blood lymphocytes	2.0	17,844	11,910	66.7%	7202	4407	1147	264	116
Total	104.6	142,872	48,339	33.8%	18,677	53,402	29,661	7200	4270

The column labels are clones: the number of clones (in millions) sequenced for this tissue; signatures: the number of different significant signatures observed in the tissue, excluding those mapping to contaminants (ribosomal, mitochondrial, repetitive elements), those mapping to more than four genes, and those mapping with single nucleotide mismatches; mapped: the number of signatures reliably mapped to transcripts (plus strand of known exons); % mapped: the percentage of signatures that were reliably mapped; genes: the number of transcribed regions to which the reliable signatures map; in loci: the number of signatures that map within loci, but to regions not known to be transcribed (in introns or within 5 kb downstream of polyadenylation site); reverse: the number of signatures that map to the reverse strand of known transcripts; genome: the number of signatures that map to the human genome outside of known loci; unmapped: the number of signatures that do not map to the human genome. Note that some categories of signatures, for example, those derived from noncoding RNAs or mapped with single nucleotide mismatches, are not represented in this table; these were excluded from the calculated percentages.

million (tpm) in at least one sample were retained. This procedure ensures that most signatures containing sequencing errors are removed (Meyers et al. 2004). A total of 182,718 distinct signatures fulfilled these criteria. The mapping of signatures to transcripts was performed essentially as described previously (Jongeneel et al. 2003). Signatures that mapped to at least one but not more than four predicted transcripts were considered to be reliable, and unreliable signatures, as well as those mapping to known contaminants or mapping with single nucleotide mismatches, were eliminated from further analysis. We were able to assign 33.8% of the different remaining signatures (87.5% of the signature counts) to known RNAs derived from 18,677 genes (Table 1). The proportion of signatures that could be assigned to known transcripts in individual tissues varied between 49.0% in cerebellum (8183 different genes) and 71.2% in pancreas (5845 genes), reflecting the depth at which these tissues have been sampled in EST and full-length cDNA sequencing projects, as well as the relative complexity of their transcriptional programs. The average efficiency of signature mapping over all tissues was 60.3%; this is significantly more than the overall efficiency (33.8%), because most of the unmapped signatures appear to be tissue-specific. The details of the mapping process are summarized in Table 1.

The signatures that could not be assigned to known transcripts are a rich source of information about the part of the transcriptome that is not yet characterized. Of the signatures, 37.4% (6.7% of the signature counts in tpm) matched mapped loci, but in regions that are not part of mapped exons (Table 1, "in loci"; note that the signature counts for each category are not shown in the table); these could represent as many as 50,000 transcripts derived from known loci but whose structure has not yet been elucidated. Another 20.7% (2.1% of tpm) matched the complementary strand of known transcripts, and could be derived from antisense or regulatory RNAs, or from overlapping genes. Collectively, 92% of all signatures and >97% of tpm mapped to the 50% of the genome that is known to be transcribed; this indicates strongly that while the full complexity of the human transcriptome may not yet have been explored, the vast majority of the transcribed regions (genes) have been identified. Only 5.0% (0.5% of tpm) mapped to the genome, but outside areas known to be transcribed, and only 3% (2.1% of tpm) of all signatures could not be mapped to the current assem-

bly of the human genome (NCBI 34, 10 Mar 2004). Because they require experimental validation, the signatures that did not map to known transcripts (Table 1, in loci + reverse + intergenic), representing 46.4% of all unique signatures, but only 10.8% of the total count, were not taken into account for the rest of this study.

Toward a definition of the adult human transcriptome

The mapped signatures matched 18,667 genes reliably (one signature matching four genes or less), and another 3494 genes unreliably (genes matched by signatures that also matched more than three other genes). The 440 signatures matching five genes or more were excluded from the rest of the study. Therefore, ~20,000 genes, roughly half of the 39,437 genes currently defined by our mapping procedures, may be expressed at detectable levels in the tissues sampled. While 39,437 may be an overestimate of the number of transcribed regions in the human genome, the numbers of genes that were expressed or not were defined relative to the same data set (the public transcriptome sequence collections) (Strausberg et al. 2002), suggesting that the estimate that half of all genes are expressed at a level detectable by this technique is reasonable. Other surveys of gene expression in normal human tissues reached similar conclusions (Hsiao et al. 2001; Su et al. 2002; Shmueli et al. 2003). As an independent verification of this estimate, we counted the number of genes from Chromosome 21 whose expression is documented by MPSS signatures, because this chromosome has undergone careful annotation. Out of 228 Chromosome 21 genes in the current Ensembl annotation, the expression of at least 126 (55%) was detected in the 32 human tissues, providing independent confirmation that about half of all genes are detectably expressed in these samples.

In a complementary approach, we examined the relationship between predicted and observed signatures (Table 2). In this context, predicted signatures are all sequences proximal to a 3′-most DpnII site in our reconstituted human transcriptome (Iseli et al. 2002; Sperisen et al. 2004), including those derived from alternatively polyadenylated or spliced transcripts, while observed signatures are a subset of the predicted ones. The predicted signatures are further subdivided into categories: specific (mapping to four transcripts or less) or nonspecific, mapping at <300 nt from the transcript 3′-end or more, and overlapping or not with the poly(A) tail. The results show several interesting features:

1. The most abundant classes of predicted signatures, as expected, are specific and do not overlap the poly(A) tail.
2. Those mapping >300 nt from the mRNA 3′-end are more than threefold less abundant on average than those mapping closer to the poly(A) tail, confirming the observation that in the classic MPSS protocol there is a bias toward signatures mapping close to the poly(A) tail; however, there is only a small difference in the observed/predicted ratio (24% vs. 27%) for these two classes.
3. Overall, 27% of the predicted signatures are actually observed; this is significantly less than half of the signatures (as compared to approximately half of all genes), but not entirely surprising given the fact that our transcript reconstitution algorithm predicts all possible transcript forms, many of which may not be present in the tissues sampled.
4. As expected, the observed overpredicted ratios for nonspecific signatures, as well as their average abundance, are much higher than for specific ones.

Overall, these results are consistent with those above, indicating that approximately one-half of all human genes defined by cDNA libraries are expressed at detectable levels in the collection of tissues sampled here. In two cell lines that we analyzed previously (Jongeneel et al. 2003), >4000 genes not found in any of the fully differentiated tissues were found to be expressed; whether this is due to the fact that genes in cell lines are less tightly regulated than those in tissues, or to a better representation of transcripts in the new MPSS protocol that was used for these cell lines remains to be determined. Also, there are almost certainly transcripts that cannot be detected because their expression is below the assay's threshold. In all tissues sampled, the frequency of signature counts was still increasing at the lower end of the distribution (data not shown), suggesting that the sampling achieved in this study (library sizes from 1.3 to 6 million clones) is still short of saturation.

Table 2. Prediction and observation of MPSS signatures among the 32 tissue samples

Predicted	Observed	Ratio	Count	Avg count	Specific	<300 nt	>300 nt	Overlaps poly(A)
7	5	0.71	470	94	No	No	No	Yes
5	1	0.20	5272	5272	No	No	Yes	No
54	43	0.80	71,543	1664	No	Yes	No	No
56	41	0.73	4147	101	No	Yes	No	Yes
396	379	0.96	4,974,891	13,126	No	Yes	Yes	No
24	21	0.88	41,669	1984	No	Yes	Yes	Yes
1495	117	0.08	12,888	110	Yes	No	No	Yes
39,664	9470	0.24	1,755,764	185	Yes	No	Yes	No
4	2	0.50	277	139	Yes	No	Yes	Yes
70,936	19,289	0.27	11,703,037	607	Yes	Yes	No	No
607	196	0.32	59,769	305	Yes	Yes	No	Yes
32,020	10,186	0.32	14,088,087	1383	Yes	Yes	Yes	No
82	53	0.65	92,829	1751	Yes	Yes	Yes	Yes
145,350	39,803	0.27	32,810,643	824				

The column labels are predicted: the number of signatures predicted in this class from transcriptome reconstitution data; observed: the number of predicted signatures actually observed; ratio: the ratio of observed to predicted signatures; count: the cumulative abundance of this class of signature; avg count: the average abundance of this class of signature; specific: signature maps to four transcripts or less (yes), or more than four (no); <300 nt: signature maps within less (yes) or more (no) than 300 nt from the transcript 3′-end; >300 nt: signature maps within more (yes) or less (no) than 300 nt from the transcript 3′-end; overlaps poly(A): signature contains A nucleotides at the 3′-end derived from poly(A) tail. NB: since one signature can map to more than one transcript, the positional classes are not mutually exclusive.

Comparing the composition and complexity of tissue-specific transcriptomes

The 32 tissues analyzed differ markedly in the apparent complexity of their transcriptomes, with 5845 genes being detected in the pancreas, while 12,267 are found in the testis (Table 1). This complexity is related to the tissues' degree of specialization, and to the number of different cell types present in them. In the pancreas, much of the transcriptional output is directed toward the manufacture of a limited

repertoire of secreted enzymes; also, because of the very high abundance of those few transcripts, the less abundant ones will fall below the significance cutoff. In the testis, no abundant tissue-specific transcripts dominate the total population, which is derived from a large number of cell types of both germ-line and somatic origin. These differences can be illustrated graphically in a cumulative histogram plotting the number of ranked transcripts against their contribution to the total transcriptome (Fig. 1). Highly specialized tissues can be clearly distinguished from more "generalist" ones in such a representation. For example, the 100 most abundant transcripts (<2% of the total number) add up to ~90% of the total mRNA in pancreas, but only 20% in fetal brain or testis. To see whether similar features could also be detected in hybridization-based data, the analysis shown in Figure 1 was repeated for both MPSS and Affymetrix data (Su et al. 2004), using a selection of tissue samples and probe sets that are common to both data sets. While the overall features of the curves are similar, differences in the distribution of abundance classes are much less marked when analyzing Affymetrix-based data, presumably because the hybridization signal reaches saturation for the most abundantly expressed genes and because the normalization method used by the Affymetrix software has dampened the distribution (Supplemental Fig. 5).

The large dynamic range of the MPSS technique allows the measurement of expression levels ranging between $>10^5$ copies per cell and less than two. Thus, individual genes can show very high degrees of tissue specificity, and be classified accordingly. Gastric lipase (LIPF), for example, was found at 9218 tpm in the stomach and less than two in all other tissues. This specificity is consistent with the distribution of the corresponding ESTs (UniGene cluster Hs.523130 at http://www.ncbi.nlm.nih.gov/UniGene) as well as SAGE tags (NlaIII tag CAGTGCTTCT, at http://cgap.nci.nih.gov/SAGE/AnatomicViewer). A simple measure of specificity can be obtained by calculating

$$S = \log_2\left(\frac{E_{\max} + 1}{\sum_{i=1}^{n} E_i - E_{\max} + 1}\right),$$

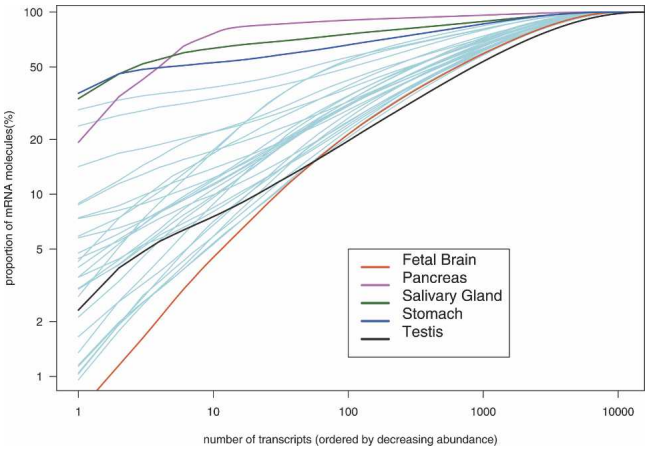

Figure 1. Distribution of transcript abundance classes in various tissues. For each tissue, the proportion of the transcriptome contributed by the *n* most abundant transcripts (abscissa) was plotted. The plots of five tissues representing extreme cases were colored: pancreas, salivary gland, and stomach as examples of highly specialized tissues with a secretory function; fetal brain and testis as examples of tissues with complex and diversified transcriptomes.

where S is the specificity, E_1 to E_n are the expression levels across all tissues, and $E_{\max}$ is the highest expression value observed for the gene in question among all tissues. Note that for this analysis, the expression levels for all adult CNS tissues were averaged into a single value. A list of the 32 genes with S values higher than 9 (i.e., expressed >512-fold higher in one tissue than in all others combined) is presented in Table 3. Most are well-known genes, whose specificity is picked up with high sensitivity by the technique. It is notable that in this set, the SymAtlas Affymetrix data document the same tissue-specific expression as the MPSS data in all cases where both tissue and probe set could be matched. The S values, however, are always significantly lower than for the MPSS data, reflecting again their narrower dynamic range (Table 3). Interestingly, there are a few highly tissue-specific genes whose identity or function remains unknown. A more comprehensive list, including all genes with an S value higher than 3 and sorted by tissue of highest expression, is given in Supplemental Table 1. There are 1759 genes in the list, of which almost half (857) are testis-specific; many known genes with an expression profile limited to germ-line cells and re-expressed in cancer (cancer-testis, or CT genes) are among the latter.

The pattern of expression of genes among tissues is also informative. Figure 2 shows that the distribution among the tissues of genes with expression values >5 tpm is bimodal, with peaks at 1 and 24 (all) tissues. This is incompatible with a model in which most or all genes would have equal probabilities to be expressed in any one tissue, which would produce a unimodal, binomial distribution. In other words, most genes are either ubiquitously expressed or tissue-specific, and their expression is not used in a primarily combinatorial fashion to produce the phenotypes of fully differentiated tissues. There are 1303 genes expressed in all samples at 5 tpm or more, giving an estimate of the number of known genes that perform "housekeeping" functions; if the threshold is increased to 10 tpm, this number falls to 942, or 2.4% of all documented genes. One should keep in mind that these numbers comprise both false positives (e.g., transcripts that are universal contaminants, such as globins), and false negatives (mostly transcripts that cannot be reliably detected by MPSS). The percentage of housekeeping genes relative to the total transcriptome (7.5%) is comparable to numbers reported by others (Warrington et al. 2000; Su et al. 2002). There are 3583 genes that are found in only one tissue, and 4403 with a specificity (as defined above) of >1, that is, expressed more than twofold higher in one tissue than in all others combined. These numbers indicate that in our collection of tissues, at least one-fifth of all expressed genes can be considered to be tissue-specific, and ~90% are not expressed in all tissues.

Tissue classification based on patterns of gene expression

Patterns of gene expression can be used to compare tissues with each other. We computed the correlation coefficient r between the logarithms of the gene expression vectors of all pairs of tissues, and used $d = (1 - r)$ as a measure of the difference between the members of a pair. The d values were used to construct a multidimensional scaling (MDS) map of the tissues (Fig. 3). The MDS method represents the 32 points in a plane while seeking to maximally preserve all the pairwise distances in the visualization. To better reveal patterns of similarity, lines connecting each sample to its nearest neighbor in the original distance matrix were added to the plot. As expected, the CNS samples generated an almost fully connected network. The retina and the pituitary

Table 3. Genes whose specificity of expression in the MPSS data (see text) was >9.0

Gene name	HGNC symbol	*S*(M)	E_{max}	*S*(A)	Organ
Elastase 3A and 3B, pancreatic	ELA3A; ELA3B	14.20	18,767	3.98	Pancreas
Lipase, gastric	LIPF	13.17	9218	NA	Stomach
Rhodopsin (opsin 2, rod pigment)	RHO	12.88	7540	NA	Retina
Pepsinogen 5, group I (pepsinogen A)	PGA5	11.68	310,995	NA	Stomach
Natriuretic peptide precursor A	NPPA	11.47	3655	0.19	Heart
Azurocidin 1 (cationic antimicrobial protein 37)	AZU1	11.18	2326	2.45	Bone marrow
Myosin, light polypeptide 2, regulatory, cardiac, slow	MYL2	11.04	2101	−0.34	Heart
Follicle-stimulating hormone, β polypeptide	FSHB	10.61	2005	−0.51	Pituitary
Cystatin SN	CST1; CST2; CST4	10.54	8950	2.72	Salivary gland
Colipase, pancreatic	CLPS	10.52	2929	4.23	Pancreas
cDNA DKFZp686M02252		10.31	1267	NA	Salivary gland
Defensin, α 6, Paneth cell-specific	DEFA6	10.27	3709	NA	Small intestine
Fascin homolog 3, actin-bundling protein, testicular	FSCN3	10.19	1169	−1.35	Testis
Kallikrein 2, prostatic	KLK2	10.18	9271	1.82	Prostate
Gastrin	GAS	10.00	1313	NA	Stomach
Trefoil factor 2 (spasmolytic protein 1)	TFF2	9.71	837	NA	Stomach
Centrin, EF-hand protein, 1	CETN1	9.65	805	−2.12	Testis
Aryl hydrocarbon receptor interacting protein-like 1	AIPL1	9.61	779	NA	Retina
Mitochondrial capsule selenoprotein	MCS; MCSP	9.57	759	1.01	Testis
Carboxypeptidase A1 (pancreatic)	CPA1	9.50	724	5.84	Pancreas
cDNA FLJ13513 fis, clone PLACE1005477		9.36	658	NA	Placenta
Pancreatic lipase-related protein 2	PNLIPRP2	9.33	8997	6.24	Pancreas
Hypothetical protein MGC42718		9.32	638	NA	Testis
Retinal G-protein-coupled receptor	RGR	9.24	865	NA	Retina
Prolactin	PRL	9.23	145,390	3.36	Pituitary
Homo sapiens testis nuclear RNA-binding protein	TENR	9.13	558	−1.10	Testis
Pancreatitis-associated protein	PAP	9.12	3897	1.83	Pancreas
Homo sapiens cDNA FLJ46224		9.11	552	NA	Testis
Transition protein 1 (histone to protamine replacement)	TNP1	9.05	6358	2.32	Testis
Myosin-binding protein C, cardiac	MYBPC3	9.05	2118	1.62	Heart
Pancreatic lipase-related protein 1	PNLIPRP1	9.02	2077	2.50	Pancreas
Testis-specific protein, Y-linked; Unknown	TSPY	9.01	516	2.02	Testis

S, the value for specificity, was calculated as described in the text. *S*(M): specificity from MPSS data; *S*(A): specificity from SymAtlas Affymetrix data. E_{max} is the maximal expression value in tpm. NA: Affymetrix data not available, either because the relevant tissue was not sampled or because there was no corresponding probe set.

gland, which are of partial CNS origin, are neighbors of CNS tissues. The three samples of hematopoietic origin (bone marrow, monocytes, and peripheral blood lymphocytes) formed a tightly connected group, as did the spleen and the thymus, which are both rich in lymphocytes. Relationships between other tissue types were more difficult to unravel in this representation.

As an alternative way to display the relationship between the gene expression profiles of different tissues, we performed a hierarchical clustering based on the same distance measure (Fig. 4). While this method does not cluster all tissues in a manner consistent with their known histological or physiological properties, several clusters (smooth muscle, intestinal tract, secretory glands, CNS) clearly emerge and are colored in the figure. A similar analysis was performed with matching subsets of the MPSS data and of Affymetrix data from the SymAtlas collection (Su et al. 2004), and the results are shown in Supplemental Figure 2. The clustering of data generated using these two very different technologies gave qualitatively similar results; in particular, the CNS samples clearly segregated from other tissues (except for fetal brain, which was separated from other CNS samples in the SymAtlas data), and the two striated muscle samples were clustered together in both data sets. The other tissues that overlap between the two data sets are too heterogeneous to cluster in a meaningful fashion.

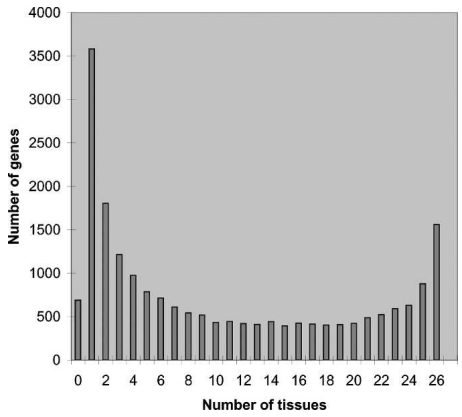

Figure 2. Frequency histogram of gene expression. For each of the genes, the tissues showing expression at 5 tpm or more were counted. The CNS samples were averaged and counted as a single tissue.

Discussion

The data presented here provide a comprehensive overview of gene expression in adult human tissues. We are making the data available for downloading, as well as providing a Web interface for interrogating them. Several other data sets documenting patterns of gene expression in normal human tissues have been published previously: Warrington et al. (2000) hybridized pooled RNA samples from 11 normal tissues (obtained from Clontech) to Affymetrix HuGeneFL chips (7129 probe sets). Hsiao et al. (2001) used the same chips to probe 59 samples derived from 19 normal

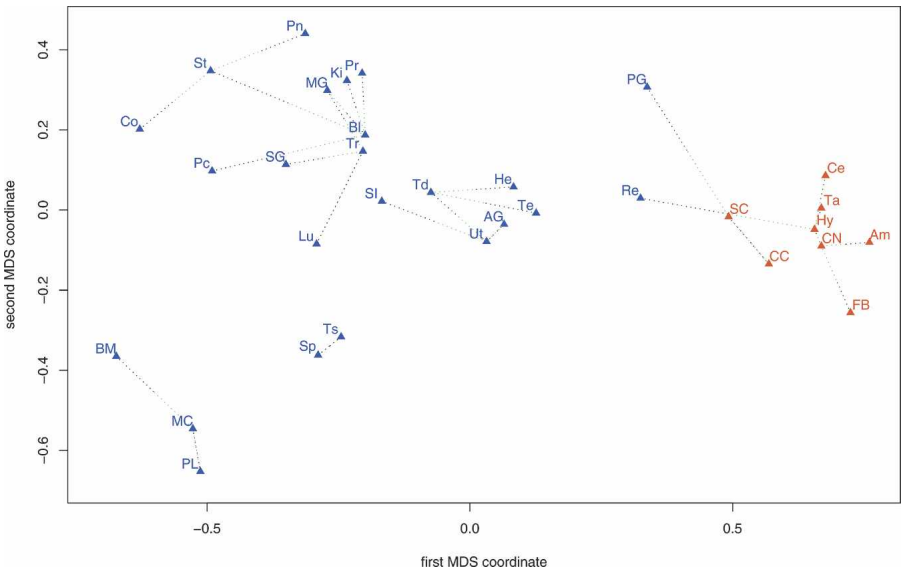

Figure 3. Multidimensional separation plot of the distance between gene expression patterns in the 32 tissues. The values of the pairwise correlations between expression vectors, *r*, were calculated from the natural logarithms of the expression values, and the distance measure $d = (1 - r)$ was used as input for the MDS routine in the multivariate analysis package of R. The CNS tissues are red. Tissues are (AG) adrenal gland; (Bl) bladder; (BM) bone marrow; (Am) amygdala; (CN) caudate nucleus; (Ce) cerebellum; (CC) corpus callosum; (FB) fetal brain; (Hy) hypothalamus; (Ta) thalamus; (He) heart; (Ki) kidney; (Lu) lung; (MG) mammary gland; (Pn) pancreas; (PG) pituitary gland; (Pc) placenta; (Pr) prostate; (Re) retina; (SG) salivary gland; (SI) small intestine; (SC) spinal cord; (Sp) spleen; (St) stomach; (Te) testis; (Ts) thymus; (Td) thyroid; (Tr) trachea; (Ut) uterus; (Co) colon; (MC) monocytes; (PL) peripheral blood lymphocytes.

human tissue types. Su et al. (2002) used the more recent HG-U95A chip (12,559 probe sets) to probe 46 samples from human tissues and cell lines. Finally, Shmueli et al. (2003) used the full HG-U95 set (62,839 probe sets on five chips) to probe 12 pooled RNA samples from human tissues also obtained from Clontech. Su et al. (2004) recently expanded their data set very significantly by designing custom probe sets that can be combined with the Affymetrix HG-U133A chip to interrogate a total of 44,775 human transcripts, and hybridizing the chips to a panel of 79 human samples. The data for four of these studies are available on the Web: http://www.hugeindex.org (Human Gene Expression Index) for Hsiao et al., http://expression.gnf.org (Gene Expression Atlas) and http://symatlas.gnf.org (SymAtlas) for Su et al., and http://genecards.weizmann.ac.il/genenote (GeneNote) for Shmueli et al. The RNA samples analyzed here, which were obtained from Clontech, overlap those used by Su et al. and by Shmueli et al.

This study differs from previously published ones in two fundamental ways: (1) signatures produced by MPSS are a largely unbiased set, derived from random sampling of polyadenylated transcripts, and the data are therefore not limited by the coverage of a given probe set; (2) MPSS allowed the measurement of 10^5-fold differences in expression, while the dynamic range of the GeneChip is $<10^3$. This difference in dynamic range was corroborated by a comparison between MPSS and GeneChip data obtained from identical mRNA samples from testis and placenta (data not shown): while 347 genes were differentially expressed by more than 100-fold according to the MPSS data, only six met the same criteria according to the GeneChip data. As an example from this data set, the pancreatic elastase 3 gene (*ELA3A/ELA3B*), which scored the highest in our specificity measurement and whose expression is known to be restricted to pancreas (Tani et al. 1988), was detected by MPSS at 18,767 tpm in the pancreas, and below the detection limit (<2 tpm) in all other tissues. In the Gene Expression Atlas, it had a score of 15,303 in the pancreas, 1324 in the spinal chord, 462 in the corpus callosum, and <100 in all other tissues, and was not flagged as a tissue-specific gene in the original paper. The GeneNote data for *ELA3A* give an expression level of ~4000 in the pancreas, 80 in the spleen, 40 in the prostate, and <20 in all other tissues sampled. In GeneNote, *ELA3A* is annotated as a tissue-specific gene. Su et al. (2002) detected 387 genes that were expressed in a tissue-specific manner, as defined by an expression level of >200 in one tissue and <100 in all others (~>10 copies per cell vs. <5). Using similar criteria, we could detect >4100. Therefore, our data set considerably extends the scope and sensitivity of available data on gene expression in adult human tissues.

There are several known limitations to the MPSS technique. The first one is that for ~7% of genes, no signatures can be reliably assigned (genes matched only by signatures marked as nonspecific in Table 2, and transcripts lacking a DpnII site), and therefore expression levels cannot be measured; this proportion doubles if one considers genes for which at least one of the possible signatures is not reliable. The use of a second anchoring enzyme, recognizing a different 4-nt sequence, should make it possible to detect those transcripts that were missed because they lack a DpnII site; however, this would significantly increase the cost of generating the data. The generation of longer signatures, which is possible at the cost of lowering the efficiency of sequencing, will also increase the proportion of reliable signatures. The second limitation is that for most genes, the detection and characterization of alternative polyadenylation sites is still fragmentary. While our transcript reconstitution procedures are able to extend many transcriptional units toward their 3′-most polyadenylation site, there are also many 3′-UTRs that are still disconnected from their parent gene and therefore not annotated properly (Iseli et al. 2002). The third limitation is associated with the original Megaclone technique (Brenner et al. 2000b), which was used in the present study. Because the entire region between the polyadenylation site and the first DpnII restriction site had to be converted to cDNA and amplified by PCR, the efficiency of detection of individual transcripts diminished with an increase in the distance separating these two sites (see Table 2). Recent improvements in the Megaclone protocol have eliminated this problem. A fourth issue is the reproducibility of the levels of gene expression measured by MPSS (Stolovitzky et al. 2005). While the reproducibility is good for genes for which signatures are detected in all samples, individual signatures may not be detected in some libraries for unknown reasons. Therefore, measurements of zero (signature not detected) have a significantly higher error rate than those with nonzero values. A comparison to SAGE data generated from similar tissues (A. Delaney, pers. comm.) shows

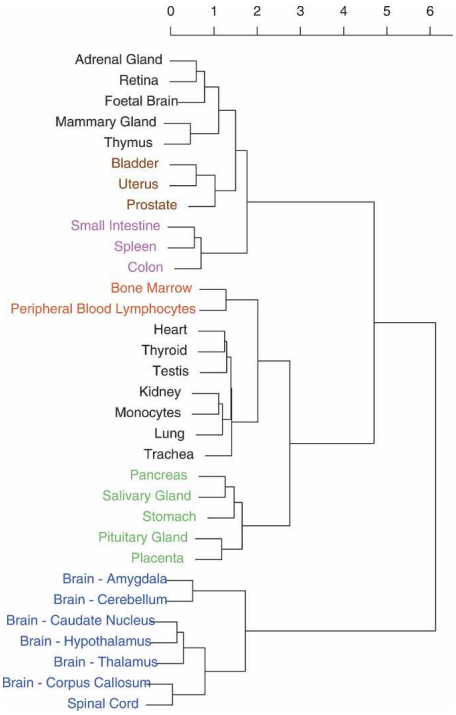

Figure 4. Hierarchical clustering of tissues based on their pairwise distances ($d = 1 - r$), using the Ward statistical method. Groups of clustered tissues are colored according to common properties: (magenta) lymphoid tissues; (red) hematopoietic tissues; (green) intestinal tract; (blue) central nervous system.

that more sequencing is required for MPSS to reach a similar coverage of the transcriptome, but that because the sampling is deeper with MPSS the quantitation is more reliable.

It has been argued that the systematic sampling of the human transcriptome using unbiased techniques such as SAGE or MPSS, or hybridization to whole genome probe sets, would uncover a vast new landscape of transcripts that had not been characterized before (Chen et al. 2002; Kapranov et al. 2002). Our data allow us to address this question directly. Only 34% of the signatures that were collected could be mapped to transcripts (Table 1), indeed suggesting that >65% of the transcriptome is yet to be characterized. But a closer look at how these 65% are distributed shows a different picture (Table 1). Almost 38% map within transcribed loci, on the strand known to be transcribed, but outside mapped exons. These could define new exons, or be derived from incompletely spliced transcripts. Another 21% map to the reverse strand of known transcripts; these could be derived from antisense transcripts, a transcript type now amply documented, or from artifacts in the cDNA cloning procedure that generates the signatures. Only 5% map to the genome outside of regions that are known to be transcribed, which themselves cover <50% of the genome; this strongly supports the argument against the intergenic regions containing significant numbers of new genes. If one considers the cumulated abundance of the signatures, those that map to known transcripts generate almost 90% of the total. Taken together, these data strongly support the contention that the vast majority of human genes expressed at >3 tpm (approximately 1 copy per cell) in the set of normal tissues examined here have now been identified, even if many remain to be mapped out in detail and the characterization of antisense transcripts is still very fragmentary (Yelin et al. 2003). It is very likely that additional exons and antisense transcripts will be discovered, but most are likely to originate from loci that have already been delineated.

The present work brings into sharp focus the highly differential expression patterns of most genes, which result in the formation of highly specialized cell and tissue types. It highlights the fact that most gene products participate in the maintenance of specialized functions, and that only a small subset are necessary to ensure the basic structural and metabolic requirements of living cells. Finally, it provides a solid foundation in the search for organ- or tissue-specific targets of therapeutic compounds of all classes.

Methods

Total RNA preparations, derived from normal human tissues and pooled from multiple donors, were purchased from Clontech. After DNase treatment and isolation of poly(A)$^+$ RNA, these samples were used to generate cDNA libraries according to the Megaclone protocol (Brenner et al. 2000b), and signatures adjacent to poly(A) proximal DpnII restriction sites were sequenced by serial cutting and ligation of decoding adapters (Brenner et al. 2000a). Each signature comprised 17 nt, including the DpnII recognition sequence (GATC). Between 1.5 and 6 million signatures were sequenced from each sample, in two reading frames offset by 2 nt. Only signatures that were seen in two independent sequencing runs, and present at a minimum of 3 tpm in at least one sample, were retained for the analysis (Meyers et al. 2004). For many signatures, counts of <3 tpm were observed in some tissues; when a particular signature was observed at one copy or not at all in a given tissue, we estimated that it was expressed below a detection threshold of 2 tpm.

The mapping of signatures to human transcripts was performed essentially as described before (Jongeneel et al. 2003), using the NCBI 34 assembly of the human genome. Additionally, sequence variants present in EST sequences but not in the genomic reference sequence were taken into account for the mapping. Two different annotated files were produced: in the "signature-centric" version, each signature was associated with one or more transcribed loci or with known mitochondrial or ribosomal transcripts and repetitive sequences, or marked as unmatched; in the "gene-centric" version, only those signatures that matched transcribed regions reliably were retained, and the corresponding counts were pooled when multiple signatures mapped to the same gene (usually through alternative polyadenylation). The gene-centric file was used to document patterns of gene expression across tissues.

Simple analyses of the results were performed using *awk* and perl scripts, or Excel functions, on the annotated files. All statistical analyses were run in the R environment, in particular with functions of the *mva* and *cluster* libraries. Clustering was performed with the hierarchical clustering algorithm *agnes* (Kauffman and Rouseeuw 1990) of the *cluster* library.

Acknowledgments

This work was supported by the Ludwig Institute for Cancer Research and by the US National Cancer Institute. M.D. was supported by the National Centre of Competence in Research (NCCR) Molecular Oncology, a research program of the Swiss National Science Foundation.

References

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., et al. 2000a. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18:** 630–634.

Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J.I., Luo, S., Kirchner, J.J., Eletr, S., et al. 2000b. In vitro cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci.* **97:** 1665–1670.

Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J.D., and Wang, S.M. 2002. Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl. Acad. Sci.* **99:** 12257–12262.

Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P., et al. 2001. A compendium of gene expression in normal human tissues. *Physiol. Genomics* **7:** 97–104.

Iseli, C., Stevenson, B.J., de Souza, S.J., Samaia, H.B., Camargo, A.A., Buetow, K.H., Strausberg, R.L., Simpson, A.J., Bucher, P., and Jongeneel, C.V. 2002. Long-range heterogeneity at the 3′ ends of human mRNAs. *Genome Res.* **12:** 1068–1074.

Jongeneel, C.V., Iseli, C., Stevenson, B.J., Riggins, G.J., Lal, A., Mackay, A., Harris, R.A., O'Hare, M.J., Neville, A.M., Simpson, A.J., et al. 2003. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc. Natl. Acad. Sci.* **100:** 4702–4705.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.

Kauffman, L. and Rouseeuw, P. 1990. *Finding groups in data*. Wiley, New York.

Meyers, B.C., Tej, S.S., Vu, T.H., Haudenschild, C.D., Agrawal, V., Edberg, S.B., Ghazal, H., and Decola, S. 2004. The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res.* **14:** 1641–1653.

Shmueli, O., Horn-Saban, S., Chalifa-Caspi, V., Shmoish, M., Ophir, R., Benjamin-Rodrig, H., Safran, M., Domany, E., and Lancet, D. 2003. GeneNote: Whole genome expression profiles in normal human tissues. *C R Biol.* **326:** 1067–1072.

Sperisen, P., Iseli, C., Pagni, M., Stevenson, B.J., Bucher, P., and Jongeneel, C.V. 2004. trome, trEST and trGEN: Databases of predicted protein sequences. *Nucleic Acids Res.* **32:** D509–D511.

Stolovitzky, G.A., Kundaje, A., Held, G.A., Duggar, K.H., Haudenschild, C.D., Zhou, D., Vasicek, T.J., Smith, K.D., Aderem, A., and Roach, J.C. 2005. Statistical analysis of MPSS measurements: Application to the study of LPS-activated macrophage gene expression. *Proc. Natl. Acad. Sci.* **102:** 1402–1407.

Strausberg, R.L., Buetow, K.H., Greenhut, S.F., Grouse, L.H., and Schaefer, C.F. 2002. The cancer genome anatomy project: Online resources to reveal the molecular signatures of cancer. *Cancer Invest.* **20:** 1038–1050.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465–4470.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci.* **101:** 6062–6067.

Tani, T., Ohsumi, J., Mita, K., and Takiguchi, Y. 1988. Identification of a novel class of elastase isozyme, human pancreatic elastase III, by cDNA and genomic gene cloning. *J. Biol. Chem.* **263:** 1231–1239.

Warrington, J.A., Nair, A., Mahadevappa, M., and Tsyganskaya, M. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* **2:** 143–147.

Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21:** 379–386.

Web site references

http://cgap.nci.nih.gov/SAGE/AnatomicViewer; SAGE.
http://expression.gnf.org; Gene Expression Atlas.
http://genecards.weizmann.ac.il/genenote; GeneNote.
http://mpss.licr.org and http://sgb.lynxgen.com; MPSS atlas of human gene expression (this paper).
http://symatlas.gnf.org; SymAtlas.
http://www.hugeindex.org; Human Gene Expression Index.
http://www.ncbi.nlm.nih.gov/UniGene; UniGene.

Received September 8, 2004; accepted in revised form April 21, 2005.

An atlas of human gene expression from massively parallel signature sequencing (MPSS)

C. Victor Jongeneel, Mauro Delorenzi, Christian Iseli, et al.

Genome Res. 2005 15: 1007-1014
Access the most recent version at doi:10.1101/gr.4041005