Defining diversity, specialization, and gene specificity in transcriptomes through information theory

Octavio Martínez*† and M. Humberto Reyes-Valdés‡

*Laboratorio Nacional de Genómica para la Biodiversidad (Langebio), Cinvestav, Campus Guanajuato, Apartado Postal 629, C.P. 36500 Irapuato, Guanajuato, Mexico; and ‡Department of Plant Breeding, Universidad Autónoma Agraria Antonio Narro, Buenavista, C.P. 25315 Saltillo, Coahuila, Mexico

Communicated by Luis Herrera-Estrella, Center for Research and Advanced Studies, Guanajuato, Mexico, April 10, 2008 (received for review March 10, 2008)

The transcriptome is a set of genes transcribed in a given tissue under specific conditions and can be characterized by a list of genes with their corresponding frequencies of transcription. Transcriptome changes can be measured by counting gene tags from mRNA libraries or by measuring light signals in DNA microarrays. In any case, it is difficult to completely comprehend the global changes that occur in the transcriptome, given that thousands of gene expression measurements are involved. We propose an approach to define and estimate the diversity and specialization of transcriptomes and gene specificity. We define transcriptome diversity as the Shannon entropy of its frequency distribution. Gene specificity is defined as the mutual information between the tissues and the corresponding transcript, allowing detection of either housekeeping or highly specific genes and clarifying the meaning of these concepts in the literature. Tissue specialization is measured by average gene specificity. We introduce the formulae using a simple example and show their application in two datasets of gene expression in human tissues. Visualization of the positions of transcriptomes in a system of diversity and specialization coordinates makes it possible to understand at a glance their interrelations, summarizing in a powerful way which transcriptomes are richer in diversity of expressed genes, or which are relatively more specialized. The framework presented enlightens the relation among transcriptomes, allowing a better understanding of their changes through the development of the organism or in response to environmental stimuli.

biological complexity | gene expression | microarrays | serial analysis of gene expression (SAGE) | Shannon entropy

The transcriptome is highly dynamic; the relative transcription frequencies of the genes change in response to environmental and internal stimuli redirecting the functional and structural landscape of living organisms. Currently, we can measure transcriptome changes by counting gene tags with technologies as serial analysis of gene expression (SAGE) (1), massively parallel signature sequencing (MPSS) (2), pyrosequencing of cDNA libraries obtained from mRNA (3), or alternatively by measuring light signals in DNA microarrays (4). In any case, it is difficult to completely understand the global changes that occur in the transcriptome, given that thousands of gene frequency measurements are involved. Here, we present a set of indexes that allow the calculation of transcriptome diversity and context specialization and the degree of gene specificity. These indexes are based on the adaptation of Shannon's information theory (5) to the transcriptome framework. Our approach is exemplified by the analysis of a dataset of the transcriptome of 32 human tissues, from which >32 million gene tags were obtained (6) and a comparable dataset for the expression of human genes in 36 human tissues using the Affymetrix GeneChip for the human genome (7). The main conclusion of our study is that this conceptualization allows elucidation of aspects of the transcriptome previously uncharacterized due to the quantity and complexity of the data.

Information theory was pioneered by Claude E. Shannon in a seminal paper in 1948 (5), and it has been generalized and applied to many scientific fields (8). In particular, it has been repeatedly applied to genetics in distinct contexts (9–12). Our approach consists of considering as symbols, in the sense of information theory, the distinct transcripts found in a tissue and counting their abundance to calculate information parameters.

Results and Discussion

Theoretical Framework. Consider the division of an organism in tissues; the transcriptomes of each tissue can then be simply described as the set of relative frequencies, p_{ij}, for the ith gene ($i = 1, 2, \ldots, g$) in the jth tissue ($j = 1, 2, \ldots, t$). Then the diversity of the transcriptome of each tissue can be quantified by an adaptation of Shannon's entropy formula,

$$H_j = -\sum_{i=1}^{g} p_{ij}\log_2(p_{ij}). \quad [1]$$

H_j will vary from zero when only one gene is transcribed up to $\log_2(g)$, where all g genes are transcribed at the same frequency: $1/g$. If we consider the average frequency of the ith gene among tissues, say,

$$p_i = \frac{1}{t}\sum_{j=1}^{t} p_{ij}, \quad [2]$$

and define gene specificity as the information that its expression provides about the identity of the source tissue as

$$S_i = \frac{1}{t}\left(\sum_{j=1}^{t}\frac{p_{ij}}{p_i}\log_2\frac{p_{ij}}{p_i}\right). \quad [3]$$

S_i will give a value of zero if the gene is transcribed at the same frequency in all tissues and a maximum value of $\log_2(t)$ if the gene is exclusively expressed in a single tissue. To quantify the tissue specialization we can obtain for each jth tissue, the average of the gene specificities, say,

$$\delta_j = \sum_{i=1}^{g} p_{ij}S_i. \quad [4]$$

δ_j varies from zero if all genes expressed in the tissue are completely unspecific ($S_i = 0$ for all i) up to a maximum of $\log_2(t)$, when all genes expressed in the tissue are not expressed

Author contributions: O.M. and M.H.R.-V. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

†To whom correspondence should be addressed. E-mail: omartine@ira.cinvestav.mx.

This article contains supporting information online at www.pnas.org/cgi/content/full/0803479105/DCSupplemental.

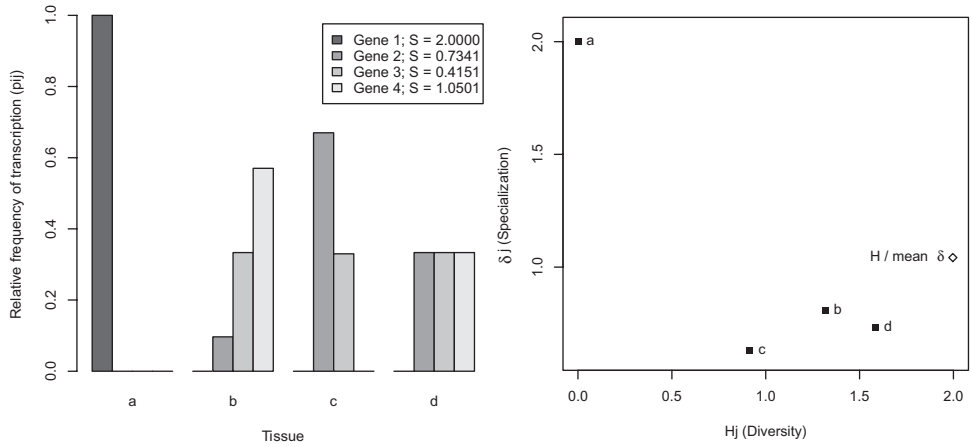

Fig. 1. Example of values of information parameters as functions of the relative frequencies of transcription of genes in a system with four tissues and four genes. (*A*) Bar plot of the relative frequencies of transcription of each gene in the four tissues. Values of gene specificities, S_i, are presented. (*B*) Scatter plot of H_j (diversity) vs. δ_j (specialization, given by the average of the gene specificities) for each tissue. The value of H in the whole system and the mean of δ_j is plotted as a diamond.

anywhere else. If we substitute the values of p_{ij} by p_i in Eq. **1** and ignore the subindex j, we obtain a measure, say H, of the diversity of the whole system.

To define a measure of divergence with respect to the whole average transcriptome, let us define the average $\log_2$ of the global transcript frequencies in a given tissue, say

$$H_{Rj} = -\sum_{i=1}^{g} p_{ij}\log_2(p_i). \qquad [5]$$

H_{Rj} will be equal to or larger than the corresponding H_j, reaching equality if and only if $p_i = p_{ij}$ for all values of i. Now we can define the Kullback–Leibler divergence of the tissue j as

$$D_j = H_{Rj} - H_j. \qquad [6]$$

D_j measures how much a given tissue j departs from the corresponding transcriptome distribution of the whole system.

Notice that H_j, the measure of diversity, depends only on the relative transcription frequencies of the tissue j; thus, it is independent of the context. However, the measures of tissue specialization and divergence, δ_j and D_j, respectively, depend not only on these frequencies but also on those of the remaining tissues; thus, these parameters are sensitive to the context where they are measured [see supporting information (SI) *Text*].

So far, we have been assuming the subdivision of an organism in a set of tissues, but the transcriptome can also be analyzed at the individual cell level or at higher hierarchic levels as sets of tissues (organs) or collections of organs (systems), etc. Transcriptome analysis can also be approached by analyzing the same organ or tissue under distinct developmental or environmental conditions. For example, we can monitor the changes in transcriptome from a normal to a malignant tumor or the effect of environmental stresses in plant transcriptomes. The framework presented here is completely general and can be used to study transcriptome changes in complex experiments.

Simple Example. Fig. 1 presents a simple and unrealistic example to illustrate the numerical results of the indexes presented here.

From Fig. 1, we can see that tissue a, which transcribes only the most specific gene ($S_1 = 2$), is the least diverse and most specialized of tissues, whereas d, which transcribes three genes at the same relative frequency, is the most diverse and the second least specialized. Tissue c, transcribing two genes with low specificities at distinct frequencies, is the least specialized with an approximate intermediate diversity, whereas tissue b, transcribing three genes, one with relatively high specificity ($S_4 = 1.05$), is the second most diverse and specialized. The diversity of the whole system, with $H = 1.9965$, almost reaches the maximum diversity for a system with four genes, $\log_2(4) = 2$, and the mean average diversity of the tissues, mean $(\delta_j) = 1.0424$, is almost in the center of the range of possible diversities, 0 to $\log_2(4) = 2$ (diamond in Fig. 1*B*). In this example, the properties of the transcriptome can be easily understood by inspection of the transcription frequencies of the four genes, but in any real case, thousands of genes are involved, and the appreciation of the transcriptome properties becomes impossible without the tools described here.

Analysis of Human Data: The Tissue Perspective. To exemplify our approach with real cases, we analyzed two comparable datasets. The first consists of >31 millions MPSS tags for 22,935 genes measured in 32 human tissues (6), and the second is a microarray expression profiling of 36 human tissues (7). These two datasets share 28 human tissues and thus present the possibility of comparing the results of our approach with two highly dissimilar methodologies.

Fig. 2 shows a scatter plot of the values of diversity, H_j vs. the values of specialization given by the average gene specificity of the tissues, δ_j.

From the results of the MPSS dataset (Fig. 2 *A* and *C*), note that the less diverse and more specialized organ is the pancreas, followed by the salivary gland and stomach. As noted in ref. 6, much of the transcriptional output in the pancreas is directed toward the manufacture of a limited repertoire of secreted enzymes and, to some extent, the same can be said about the salivary gland and the stomach. We can also note how the organs of the digestive system cover almost the entire specialization spectrum, from the highest specialization of the pancreas to the relative low specialization of the small intestine. When comparing the values of H_j for the constitutive organs of the digestive system in the MPSS dataset, we see a high degree of variation from 5.2 for the pancreas up to more than double that quantity, 10.7, for the small intestine. The organs of the CNS are scattered in a region of high diversity but relatively low specialization (Fig. 2 *B* and *D*). The testis is the organ with the most diverse transcriptome; this is reasonable, because, as noted in ref. 6, 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. Among the organs sampled from the reproductive system, the placenta is more

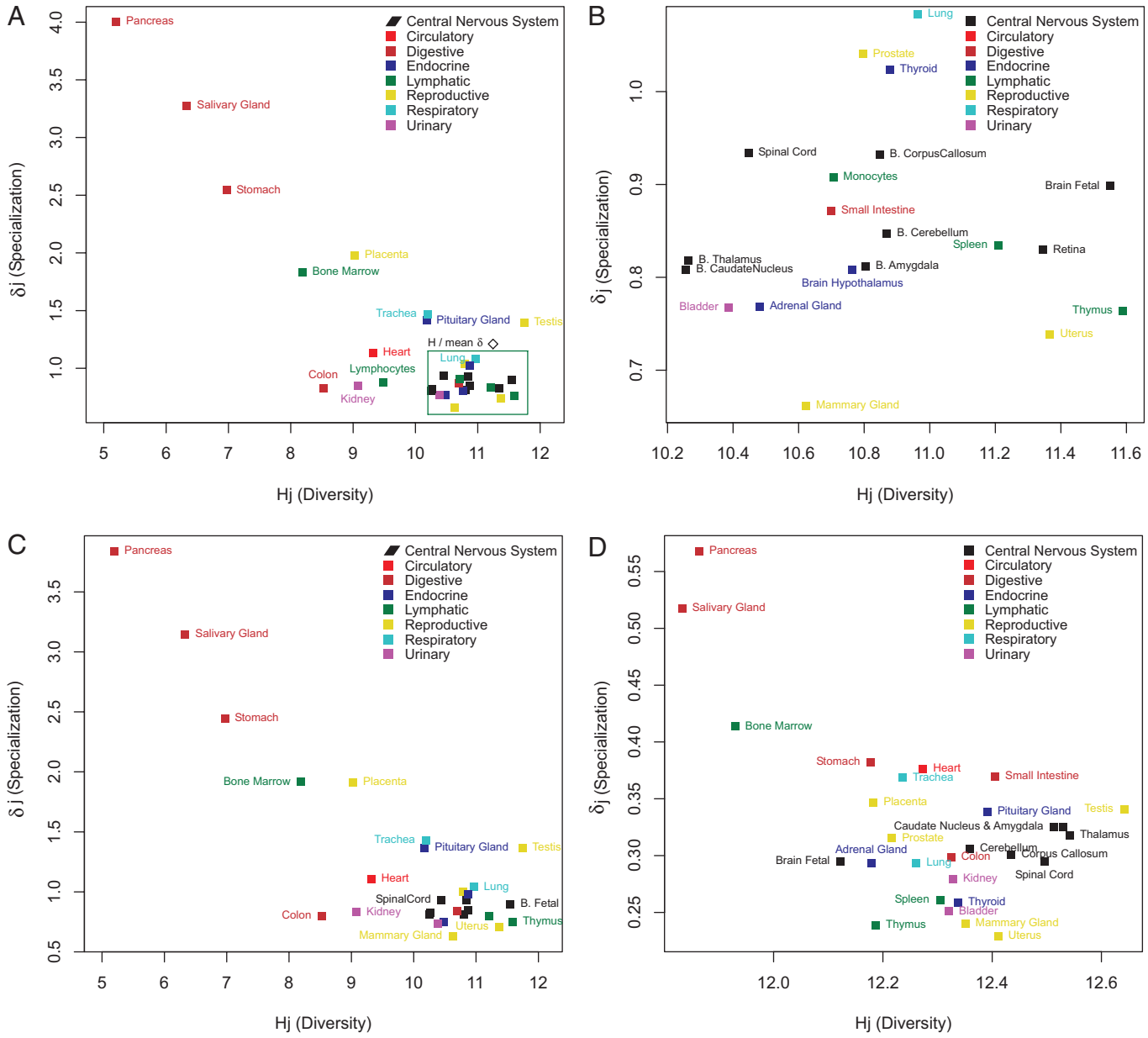

Fig. 2. Scatter plot of estimated values of H_j (diversity) vs. δ_j (specialization, given by the average gene specificity) for tissues of the human systems. Tissues are colored by system of origin. (*A* and *B*) Results from the MPSS dataset (32 tissues); *B* is an amplification of the box in *A*. The point of *H* for the whole system and the mean of δ_j is shown by a diamond in *A*. (*C*) Results from the MPSS dataset in 28 tissues shared with the microarray dataset. (*D*) Results from the microarray dataset in the shared 28 human tissues.

specialized than the testis but less diverse. Within the lymphatic system, the bone marrow is the most specialized and the least diverse. Within the endocrine system, which includes the hypothalamus given its mainly hormonal role (MPSS dataset only, Fig. 2 *A* and *B*), the pituitary gland is the most specialized and diverse. For the two organs representing the respiratory system, lung, and trachea, the lung presents a more diverse and less specialized transcriptome than the trachea, whereas for the two organs representing the urinary system, the kidney has a more specialized transcriptome than the bladder. All these observations are consistent in both datasets.

When comparing the analyses resulting from the two distinct datasets (Fig. 2 *C* and *D*), we notice a difference in the ranks covered by H_j and δ_j in the two images. The ranks of H_j and δ_j are narrower when estimated from the microarrays compared with the estimation from the MPSS data, because fewer genes with less average variation are represented in the microarray compared with the MPSS dataset. Despite these scale differences, scatter plots *C* and *D* (Fig. 2 *C* and *D*) are remarkably similar, taking into account that they arose from two completely distinct methods and used different biological samples that surely present individual noise in the estimation of the gene frequencies. In both graphs, the most specialized tissue is the pancreas, followed by the salivary gland, and the most diverse is the testis. The Pearson's correlation between the paired estimates of H_j from both datasets was $r = 0.68$, whereas the corresponding coefficient for δ_j was $r = 0.90$. Figs. S1, S2, and S3 show scatter plots for the values of H_j, δ_j, and D_j, respectively, estimated from each dataset. Fig. S4 shows the scatter plot of H_j vs. δ_j in the microarray dataset, including all 36 human tissues.

The visualization of the positions of transcriptomes in a system of diversity and specialization coordinates, as the one presented in Fig. 2 and Fig. S4, permits a full and comprehensible appreciation of these transcriptome properties that is unfeasible by other means. In ref. 6, the authors show the distribution of transcript abundance classes in various tissues, plotting the proportion of the transcriptome contributed by the n-most abundant transcripts. Using this approach, they conclude that the pancreas, salivary gland, and stomach are examples of highly specialized tissues, whereas the fetal brain and testis are presented as examples of tissues with complex and diversified transcriptomes. More subtle and detailed conclusions are reached by observing Fig. 2, which presents a complete and easy-to-interpret panorama of the diversity and specialization in all sampled tissues.

Fig. S5 presents a scatter plot of estimated values of D_j (divergence) vs. δ_j (specialization) for tissues of the human systems resulting from the MPSS dataset. In Fig. S5, we can appreciate distinct strategies of specialization of the tissue's transcriptomes. Tissues with $\delta_j > D_j$ are above the red line that marks $D_j = \delta_j$, whereas tissues with $\delta_j < D_j$ are below that line. Tissues with $\delta_j > D_j$ have a specialization strategy that consists mainly of expressing highly specialized genes, whereas tissues with $\delta_j < D_j$ achieve their specialization by expressing at higher or lower rates genes that are, on average, expressed in the whole system. The distance of each point (tissue) to the line $D_j = \delta_j$ denotes how extreme is the specialization strategy. From Fig. S5, we notice, for example, that the tissues of the reproductive system are in general very close to the line $D_j = \delta_j$, indicating an almost neutral specialization strategy. In contrast, all of the tissues of the digestive system have large deviations from the neutral specialization strategy and all, except the colon, have values of $\delta_j > D_j$, denoting a specialization strategy consisting of the expression of mainly specialized genes. The plotting of D_j (divergence) vs. δ_j (specialization), as the one shown in Fig. S5, offers immediate and easy-to-interpret insights into the specialization strategies of the transcriptomes, which will be very difficult, if not impossible, to attain without the information tools presented. Fig. S6 presents the scatter plot of H_j vs. δ_j in the microarray dataset, including the 28 tissues shared with the MPSS dataset.

In the case of the human dataset, the information analysis can also be performed at system level by grouping the sets of tissues into their corresponding systems. Fig. S7 presents the graphical result of that analysis performed over the MPSS dataset, showing a scatter plot of estimated values of H_j (diversity) vs. δ_j (specialization) that also includes the estimated values of D_j (divergence) for each system. These results are consistent with the analysis performed at the finer level of tissues presented earlier.

Analysis of Human Data: The Gene Perspective. The values of S_i calculated for each of the 22,935 genes studied in the MPSS dataset allow the quantitative classification of gene specificity, a concept regularly used in the literature but seldom quantified (13). In this case, the maximum value of S_i for the human genes, $S_i = \log_2(32) = 5$, reached by 2,555 genes (11.14%), indicates that the gene is exclusively transcribed in only one of the 32 tissues studied, whereas the minimum possible value of S_i, zero, unattained in the MPSS dataset, would indicate a gene with exactly the same frequency in all tissues. Housekeeping genes have small values of S_i, indicating an even distribution across tissues. An index for gene specificity that depends basically on its maximum expression was applied to the human dataset (6). That index does not have definite maximum or minimum bounds and misses many exclusive genes, giving values that, in contrast with our index S_i, depend not only on the specificity of the gene but also on its frequency of expression. Although the index S_i easily selects all 2,555 specific genes, the index proposed in ref. 6 can induce a misleading inference about specificity, because many specific genes with a value of $S_i = 5$ give very low values of the index proposed by ref. 6 and thus will not be classified as specific with that index. Fig. S8 presents a scatter plot for the values of both coefficients. Despite these differences, all genes presented in table 3 of ref. 6 with a value of their coefficient >9 also have a high value of S_i that ranks from 4.99 to 5. Table 1 presents examples of genes with extreme values of S_i and the S_i values attained in this dataset by some genes classified as housekeeping in the literature.

Table 1. Examples of genes with distinct value of specificity (S_i)

Completely specific (S_i = 5), highly expressed genes (tissue)		
S_i	HUGO	Description (tissue where expressed)
5.00	*LIPF*	Lipase (stomach)
5.00	*ELA3B*	Enastase 3B (pancreas)
5.00	*RHO*	Rhodopsin; opsin 2, rod pigment (retina)
5.00	*AZU1*	Azurocidin 1 (bone marrow)
5.00	*MYL2*	Myosin, light polypeptide 2 (heart)
Genes with the lowest values of S_i (expressed in all sampled tissues)		
S_i	HUGO	Description
0.09	*PSMB6*	Prosome, macropain; subunit, beta type, 6
0.09	*CHMP4A*	Chromatin modifying protein 4A
0.09	*COMMD3*	B lymphoma Mo-MLV insertion region
0.10	*CTNND1*	Catenin (cadherin-associated protein), delta 1
0.10	*PSMC5*	Prosome, macropain; 26S subunit, ATPase, 5
Genes reported as housekeeping in the literature		
S_i	HUGO	Description
0.22	*PPIA*	Cyclophilin A
0.25	*ACG1*	Actin, gamma 1
0.28	*PGK1*	Phosphoglycerate kinase 1
0.28	*TAF11*	TAF11; TATA box-binding protein
2.30	*GAPDH*	Glyceraldehyde-3-phosphate dehydrogenase

HUGO, Human Genome Organization.

Table 1 presents five examples of the 2,555 completely specific genes ($S_i = 5$) selected to be shown from the MPSS dataset, because they are the ones with the five highest average expression levels (highest p_i) and are also presented as examples in ref. 6. As mentioned above, no gene attached the minimum possible value of $S_i = 0$ that will indicate exactly the same expression level of transcription in all 32 tissues; however, Table 1 presents the five genes with the lowest values of S_i that rank from 0.09 to 0.10 and can be classified as housekeeping, because they present the most even distribution of transcription expression among the human tissues sampled. Two genes with the lowest S_i in Table 1, *PSMB6* and *PSMC5*, belong to the proteasome that is responsible for the degradation of abnormal intracellular proteins (14). The gene *CHMP4A*, which has the second smallest value of S_i, 0.09 (Table 1), is a member of the family of small coiled-coil proteins named *CHMP* implicated in playing roles in multivesicular body sorting (15), whereas *COMMD3*, with a value of $S_i = 0.09$, is a member of a gene family defined by the presence of a conserved and unique motif termed the COMM (copper metabolism gene *MURR1*) domain, which functions as an interface for protein–protein interactions. In particular, *COMMD3* has been independently shown to be expressed at relatively even levels in 13 human tissues (16). The *CTNND1* gene with a value of $S_i = 0.10$ corresponds to catenin, a protein linked to the cytoplasmic domain of transmembrane cadherins (17). These examples show that measuring the specificity of genes by S_i can lead to the detection of new housekeeping genes.

Table 1 also presents the values of S_i for genes repeatedly reported in the literature as housekeeping for human studies (18). Four of these genes (*PPIA*, *ACG1*, *PGK1*, and *TAF11*) have values of S_i between 0.22 and 0.28 that are more than double the

values for genes with the smallest values of S_i but can still be considered to have an even distribution and thus housekeeping genes. In contrast, the gene for GAPDH, the most popular housekeeping gene (18), presents a value of $S_i = 2.3$ that is almost in the center of the possible rank of S_i (0–5) and cannot be considered as housekeeping, at least for the tissues studied.

Fig. S9 presents bar plots for the distribution of 10 specific genes ($S_i = 5$; Fig. S9*A*) and the 10 genes with the lowest values of S_i (Fig. S9*B*) from the MPSS dataset, where one can appreciate how specific genes are expressed in only one organ, whereas nonspecific or housekeeping genes have an approximately even distribution among tissues.

Comparing the distributions of S_i in the systems and tissues (Fig. S10), one can appreciate in both cases an approximate U-shaped distribution, with a larger number of genes having values closer to the limits of the S_i rank. The largest difference between the distributions of S_i is observed in the first class, which in both cases groups the first fifth of the S_i rank, and that for the system distribution represents 36% of the genes, whereas for tissues, it groups 25% of the genes. This shows that, when grouping the data by systems, more genes can be classified as housekeeping or ubiquitously distributed among the systems than when grouping by tissues. The difference in the last class of the distributions, grouping one-fifth of the most extreme or specialized genes, is only of ≈3%, showing that this class of genes is less affected in the relative S_i value by the grouping than genes with a low value of S_i.

General Considerations. The transcriptome is vastly dynamic; frequencies of gene expression in tissues change during the development of the organism and at the same developmental stage in response to internal or external stimuli, modifying the landscape of the proteome and the functional and structural roles of the cells. In many instances in the recent literature on transcriptomes (19–23), the concept of complexity is mentioned in relation to the number of genes expressed and the changes of expression patterns in distinct situations; however, problems of quantitative evaluation of the transcriptome diversity or specialization and gene specificity are not addressed. The analytical tools herein presented (H_j for measuring diversity, δ_j for assessing context specialization, D_j for transcriptome divergence, and S_i for estimating gene specificity) allow the understanding of these global changes, giving insights about the complex changes occurring during these phenomena. A decrease in H_j will indicate that fewer genes are being transcribed, or that the transcription frequencies are less uniform, whereas an increase of δ_j will signal that, on average, more specific genes are transcribed, and an increase of D_j indicates departures from the average transcriptome. With the help of the S_i index, it is possible to detect genes specific to a giving condition or that, on the contrary, are maintained approximately at the same rate of transcription under different situations.

In the examples presented here, we consider tissues of a given organism; however, the information framework presented is general, and we can speak about "subsystems" of a given organism, where each subsystem can represent an order of morphological classification, i.e., individual cells, tissues, organs, systems, etc.; a state of development, for example plantlet, flowering plant, senescent plant; normal or malignant tissue, etc.; a particular experimental treatment, such as "optimal condition" vs. "stress condition" in a model organism; and so forth.

There are statistical issues about the estimation of the information properties not detailed here. A goodness-of-fit statistic can be readily obtained by transforming the Kullback–Leibler distance D_j to test the null hypothesis that the transcriptome of a given tissue is statistically equal to a given distribution (24). Another issue is the estimation of confidence intervals for H_j, D_j, δ_j, and S_i that can be obtained by the bootstrap method and will be presented elsewhere. Another important statistical issue related to the information parameters estimation is the sample size or deepness of sampling of the transcriptome. Because many genes are transcribed at very low frequencies, small sample sizes, usually used in EST studies, are likely to miss many low-expressed genes, probably underestimating the value of H_j and distorting the true values of D_j and δ_j.

When we have a snapshot of the relative frequencies of the transcribed genes, as in the case of SAGE, MPSS, or microarray experiments, the estimation of H_j makes it possible to objectively quantify the diversity of a transcriptome, capturing this aspect of its complexity. Because H_j depends only on the relative frequencies of the expressed genes, it can be used to compare transcription diversity not only between subsystems of the same organism but also between transcriptomes of various distinct organisms, allowing comparison among taxa.

The index S_i, defined as the specificity of a gene, permits the quantification of the relative spreadness of the genes across subsystems, giving a quantitative base to define concepts such as housekeeping or specialized genes recurrently used in the literature, in many cases without a quantitative assessment of their degree of variability (13).

We have shown the method in the framework of protein-codifying genes; however, it is applicable to any kind of transcript tag available, including the precursors or mature forms of noncoding RNA as iRNA, sRNA, and so forth, and to collections of anonymous tags from tissues in an organism for which nongenome sequences are available.

Materials and Methods

The MPSS dataset, consisting of >31 millions tags for 22,935 genes measured in 32 human tissues analyzed and published (6), was kindly made available to us by Jongeneel *et al*. The data consist of the number of tags obtained for each gene in each tissue. To obtain the relative frequencies of expression of each gene at each tissue, p_{ij}, the original number of tags per gene was divided by the corresponding total number of tags in the tissue.

For the analysis of the MPSS dataset at the system level, the data from the 32 tissues were grouped into eight systems in accordance with the main functional classification of the tissues (Table S1). To obtain the relative frequencies of expression of a given gene in a specific system, we took the average of the relative frequencies of expression of that gene in the organs considered to form part of the corresponding system. From this matrix of relative frequencies, $\{p_{ij}\}$, all information parameters were calculated. The number of gene tags and the information parameters calculated from the data are presented in Table S1.

The dataset of microarray experiments used the Affymetrix GeneChip (Gene Expression Omnibus accession no. GDS1096), downloaded from the National Center for Biotechnology Information site (www.ncbi.nlm.nih.gov). The file containing the normalized measurements for all identifiers in all tissues was processed to obtain the average expression per gene in each tissue. To obtain the relative frequencies of expression of each gene at each tissue, p_{ij}, the estimated average expression of each gene in a given tissue was divided by the sum of the average expression of all genes in that tissue. Two analyses were performed, one including only the 28 tissues shared with the MPSS dataset (Fig. 2*D*) and the other including all 36 tissues (Fig. S4).

The analyses were performed within the R statistical language (25). The program designed for the analyses and the full table of results for all analyses are available on request. The program will also be deposited to form part of the R Bioconductor software.

Note Added in Proof. The value of gene specificity (Eq. **3**) is a linear function of the entropy of a gene's expression distribution applied in (ref. 26) to evaluate tissue specificity.

ACKNOWLEDGMENTS. We are grateful to Luis Herrera-Estrella and two anonymous referees for suggestions and critical review of the manuscript and to Lourdes Martínez de la Vega for help with the classification of human organs into systems. We are also grateful to Victor Jongeneel *et al*. (6), who kindly sent their original data files to us. We acknowledge financial support from Conacyt, Concyteg, Cinvestav, and Universidad Autónoma Agraria Antonio Narro.

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. *Science* 270:368–369.
2. Brenner S, *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18:630–634.
3. Agaton C, *et al.* (2002) Gene expression analysis by signature pyrosequencing. *Gene* 289:31–39.
4. Meyers BC, Galbraith DW, Nelson T, Agrawal V (2004) Methods for transcriptional profiling in plants. Be fruitful and replicate. *Plant Physiol* 135:637–652.
5. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423.
6. Jongeneel CV, *et al.* (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* 15:1007–1014.
7. Ge X, *et al.* (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* 86:127–141.
8. Taneja IJ (2001) *Generalized Information Measures and Their Applications* (Departamento de Matemática, Universidade Federal de Santa Caterina, Florianópolis, Brazil).
9. Román-Roldán R, Bernaola-Galván P, Oliver JL (1996) Application of information theory to DNA sequence analysis: A review. *Pattern Recog* 29:1187–1194.
10. Schneider TD (1997) Information content of individual genetic sequences. *J Theor Biol* 189:427–441.
11. Reyes-Valdés MH, Williams CG (2005) An entropy-based measure of founder informativeness. *Genet Res* 85:81–88.
12. Chao A, Shen T-J (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample *Environ Ecol Stat* 10:429-443.
13. Thellin O, *et al.* (1999) Housekeeping genes as internal standards: Use and limits. *J Biotechnol* 75:291–295.
14. Kwak M-K, Kensler TW (2006) Induction of 26S proteasome subunit PSMB5 by the bifunctional inducer 3-methylcholanthrene through the Nrf2-ARE, but not the AhR/Arnt-XRE, pathway. *Biochem Biophys Res Commun* 345:1350–1357.
15. Katoh K, Shibata H, Hatta K, Maki M (2004) CHMP4b is a major binding partner of the ALG-2-interacting protein Alix among the three CHMP4 isoforms. *Arch Biochem Biophys* 421:159–165.
16. Burstein E, *et al.* (2005) COMMD proteins, a novel family of structural and functional homologs of MURR1. *J Biol Chem* 280:22222–22232.
17. Keirsebilck A, *et al.* (1998) Molecular cloning of the human p120ctnCatenin gene (CTNND1): Expression of multiple alternatively spliced isoforms. *Genomics* 50:129–146.
18. Tricarico C, *et al.* (2002) Quantitative real-time reverse transcription polymerase chain reaction: Normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. *Anal Biochem* 309:293–300.
19. Frith MC, Pheasant M, Mattick JS (2005) Genomics: The amazing complexity of the human transcriptome. *Eur J Hum Genet* 13:894–897.
20. Kapranov P, *et al.* (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* 15:987–997.
21. Gustincich S, *et al.* (2006) The complexity of the mammalian transcriptome. *J Physiol* 575:321–332.
22. Dix TI (2007) Comparative analysis of long DNA sequences by per element information content using different contexts. *BMC Bioinformatics* 8:S10.
23. Sayyed-Ahmad A (2007) Transcriptional regulatory network refinement and quantification through kinetic modeling, gene expression microarray data and information theory. *BMC Bioinformatics* 8:20.
24. Senoglu B, Surucu B (2004) Goodness-of-fit tests based on Kullback-Leibler information. *IEEE Tran Rel* 53:357–361.
25. R-Development Core Team (2005) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
26. Schug J, *et al.* (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* 6:R33.