Supporting Information

Martínez and Reyes-Valdés 10.1073/pnas.0803479105

SI Text

Mutual Information I (X|Y). In all cases, the units of the herein proposed parameters are bits; however, for simplicity we will omit the units in this appendix.

Let X be a random variable that takes on values $j = 1, 2, .., t$ (subsystems: tissues or conditions), and Y a random variable that takes on values $i = 1, 2.,., g$ (genes). Since p_{ij} is the frequency of the ith transcript within the jth subsystem, $p(i|j) = p_{ij}$. Now, $p(i, j) = (1/t)p_{ij}$, given that we consider all subsystem to be equiprobably distributed; for the same reason, $p(i) = p_i$, the average frequency across subsystems.

The average mutual information $I(X|Y)$ equals by symmetry $I(Y|X)$, which in turn can be defined as

$$\begin{aligned} I(Y|X) &= H(Y) - H(Y|X) \\ &= -\sum_{i=1}^{g}\sum_{j=1}^{t} p(i,j)\log_2[p(i)] \\ &\quad + \sum_{i=1}^{g}\sum_{j=1}^{t} p(i,j)\log_2[p(i|j)] \\ &= -\sum_{i=1}^{g}\sum_{j=1}^{t} \frac{1}{t}p_{ij}\log_2(p_i) + \sum_{i=1}^{g}\sum_{j=1}^{t} \frac{1}{t}p_{ij}\log_2(p_{ij}) \end{aligned}$$

Information Meaning of Gene Specificity S_i. The measure S_i is the conditional mutual information $I(X|Y = i)$, as it will be proved with the following arguments.

The expression for Eq. **3** can be rearranged as:

$$S_i = \sum_{j=1}^{t} \frac{p_{ij}}{tp_i}\log_2\left(\frac{p_{ij}}{p_i}\right)$$

From Eq. **2**, it is easy to see that

$$\sum_{j=1}^{t} \frac{p_{ij}}{tp_i} = 1$$

This allows the following rearrangement of the expression for S_i

$$\begin{aligned} S_i &= \log_2(t) - \sum_{j=1}^{t} \frac{p_{ij}}{tp_i}\log_2(t) + \sum_{j=1}^{t} \frac{p_{ij}}{tp_i}\log_2\left(\frac{p_{ij}}{p_i}\right) \\ &= \log_2(t) + \sum_{j=1}^{t} \frac{p_{ij}}{tp_i}\left[\log_2\left(\frac{p_{ij}}{p_i}\right) - \log_2(t)\right] \\ &= \log_2(t) + \sum_{j=1}^{t} \frac{p_{ij}}{tp_i}\log_2\left(\frac{p_{ij}}{tp_i}\right) \end{aligned}$$

The negative of the second term of the last expression is the entropy $H(X|Y = i)$, i.e., the conditional uncertainty of X given $Y = i$. This is because $p(j|i) = p_{ij}/ tp_i$, as proved by the Bayes theorem:

$$p(j|i) = \frac{p(i|j)p(j)}{p(i)} = \frac{p_{ij}\frac{1}{t}}{p_i} = \frac{p_{ij}}{tp_i}$$

Thus:

$$S_i = H(X) - H(X|Y = i) = I(X|Y = i)$$

Then, the information $I(X|Y)$ is the weighted average of the S_i values across all genes:

$$I(X|Y) = \sum_{i=1}^{g} p_i I(X|Y = i) = \sum_{i=1}^{g} p_i S_i$$

Range of S_i. Given that the gene specificity S_i equals $I(X|Y = i)$, its lower and upper bounds are 0 and $log_2(t)$, respectively. From Eq. **3**, it is easy to see that the minimum is attained when $p_{ij} = p_i$ for all j, i.e., when the transcription frequency is totally uniform across subsystems. The maximum value is reached when $p_{ij} = tp_i$ for a given j, and $p_{ij} = 0$ for all others (Eq. **3**), i.e., when the gene is exclusively transcribed in one subsystem.

Range of δ_j. The maximum value for S_i has been established to be $\log_2(t)$, and this sets the upper bound of δ_j (Eq. 4), the weighted average of S_i, in $\log_2(t)$. This maximum is reached when for all i such that $p_{ij} > 0$, $p_i = p_{ij}/t$, the condition that makes $S_i = \log_2(t)$ (Eq. 3). This is the situation when all transcribed genes in the subsystem j, i.e., those with $p_{ij} > 0$, are transcribed only in that subsystem. On the other hand, $\text{Min}(S_i) = 0$, which sets the lower bound of δ_j in 0 when $p_{ij} = p_i$ for all i, j, i.e., when all genes have the same average specificity. An important feature of δ_j is that it takes a value of 0 if and only if all other subsystems have also values of 0.

The relationship between the average of δ_j and the mutual information $I(X|Y)$ can be seen in the following equality:

$$\frac{1}{t}\sum_{j=1}^{t}\delta_j = \frac{1}{t}\sum_{j=1}^{t}\sum_{i=1}^{g} p_{ij}S_i = \frac{1}{t}\sum_{i=1}^{g} S_i \sum_{j=1}^{t} p_{ij} = \sum_{i=1}^{g} p_i S_i = I(X|Y)$$

The Kullback–Leibler Divergence. This measure, based on the concept of relative entropy (1, 2), can be used to compare a distribution or series of distributions against one with fixed parameters. In this case, we use the Kullback–Leibler divergence to evaluate the departure from the transcription distribution of a given subsystem from a distribution formed by the arithmetic averages of the transcription frequencies across subsystems.

$$D_j = \sum_{i=1}^{g} p_{ij}\log_2\left(\frac{p_{ij}}{p_i}\right)$$

D_j is the weighted average of $log_2(p_{ij}/p_i)$ for the jth subsystem. Given that the log function is monotonically increasing for any base $b \geq 1$, the maximum term for any fixed p_{ij} in the log_2 function will correspond to $\text{Max}[p_{ij}/p_i]$. This maximum is attained for $p_{ij} > 0$ when $p_i = p_{ij}/t$, i.e., when all other subsystems have a transcription frequency of 0 for the ith gene; meanwhile the terms in the summation with $p_{ij} = 0$ are all 0. Thus the maximum weighted value of $log_2(p_{ij}/p_i)$ is $\log_2(t)$, which sets up an upper bound for the average D_j. This upper bound is attainable if and

only if all genes expressed in the jth subsystem are not expressed in any other subsystem, i.e., if $p_i = p_{ij}/t$ for all $p_{ij} > 0$. On the other hand, the lower bound of 0 for the Kullback–Leibler divergence is well established; thus, the range of D_j is $(0, \log_2(t))$. The minimum value is reached when $p_{ij} = p_i$ for all i in the jth subsystem. This contrasts with the case of δ_j, whose lower bound requires $p_{ij} = p_i$ for all i and all j.

The Kullback–Leibler distance D_j is related in this context with the mutual information $I(X|Y)$:

$$I(X|Y) = \frac{1}{t}\sum_{j=1}^{t}\left[\sum_{i=1}^{g} p_{ij}\log_2(p_{ij}) - \sum_{i=1}^{g} p_{ij}\log_2(p_i)\right]$$

$$= \frac{1}{t}\sum_{j=1}^{t}\sum_{i=1}^{g} p_{ij}\log_2\left(\frac{p_{ij}}{p_i}\right) = \frac{1}{t}\sum_{j=1}^{t} D_j$$

1. Ewens WJ, Grant GR (2001) *Statistical Methods in Bioinformatics* (Springer, New York), p 45.
2. Taneja IJ (2001) *Generalized Information Measures and Their Applications* (Departamento de Matemática, Universidade Federal de Santa Caterina, Florianópolis, SC, Brazil).

Thus, the average mutual information $I(X|Y)$ is the average of the D_j values across subsystems, and the following equality holds

$$I(X|Y) = I(Y|X) = \frac{1}{t}\sum_{j=1}^{t} D_j = \frac{1}{t}\sum_{j=1}^{t} \delta_j = \sum_{i=1}^{g} p_i S_i$$

From this equality, it can be seen that, regardless of the upper bounds of the individual values of D_j, δ_j and S_i, their averages will be bounded above by $\mathrm{Min}\{\mathrm{Max}[I(X|Y)], \mathrm{Max}[I(Y|X)]\}$, i.e., $\mathrm{Min}[\log_2(t), \log_2(g)]$. Furthermore, given that the first term in $I(X|Y)$ is exactly the entropy $H(X)$, already defined as the whole system diversity H, the averages of the three parameters D_j, δ_j and S_i are limited above by the entropy of the average transcription frequencies. (*S20*) gives a comprehensive presentation of the information measures.

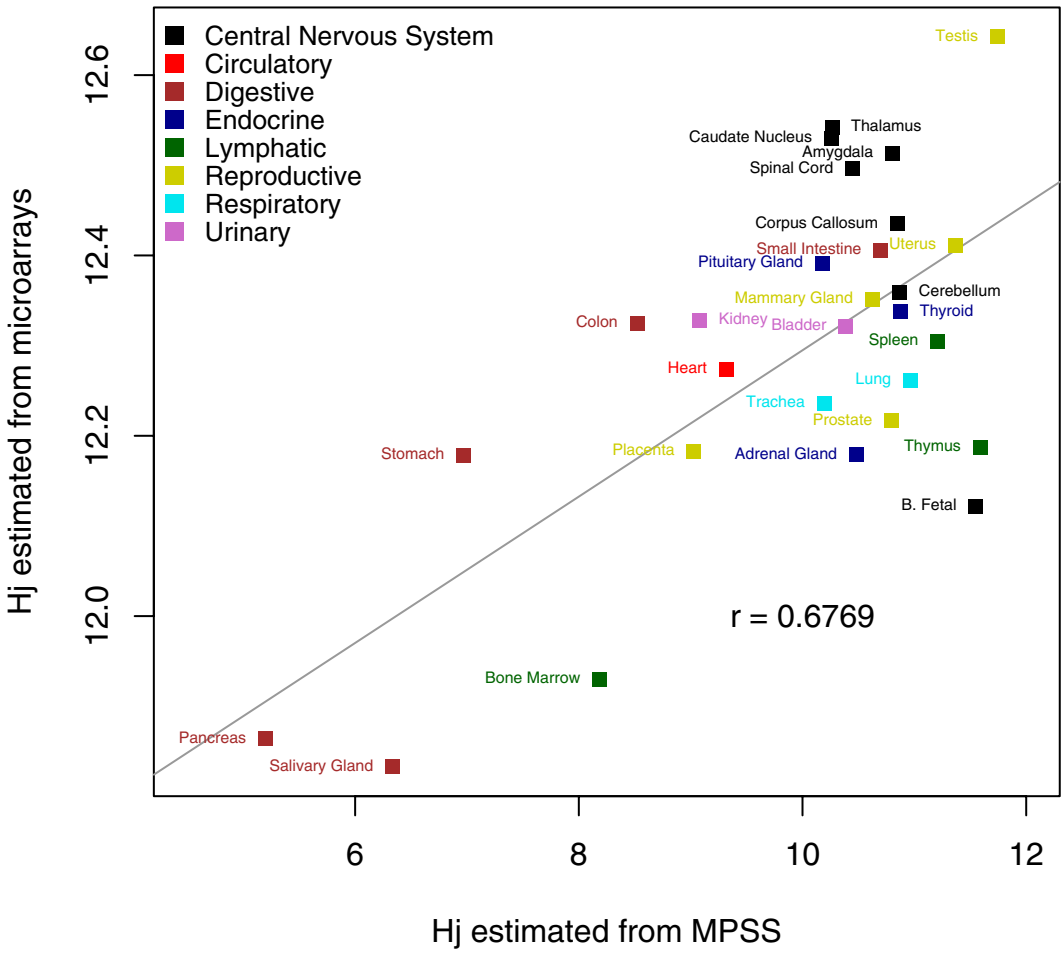

Fig. S1. Scatter plot for the estimated values of H_j from the MPSS (*x* axis) and microarray dataset (*y* axis) in the 28 shared tissues between datasets. Tissues are colored by system of origin. The gray line indicates a lineal model fitted between the two variables. The value of the r Pearson's correlation coefficient is presented as text.

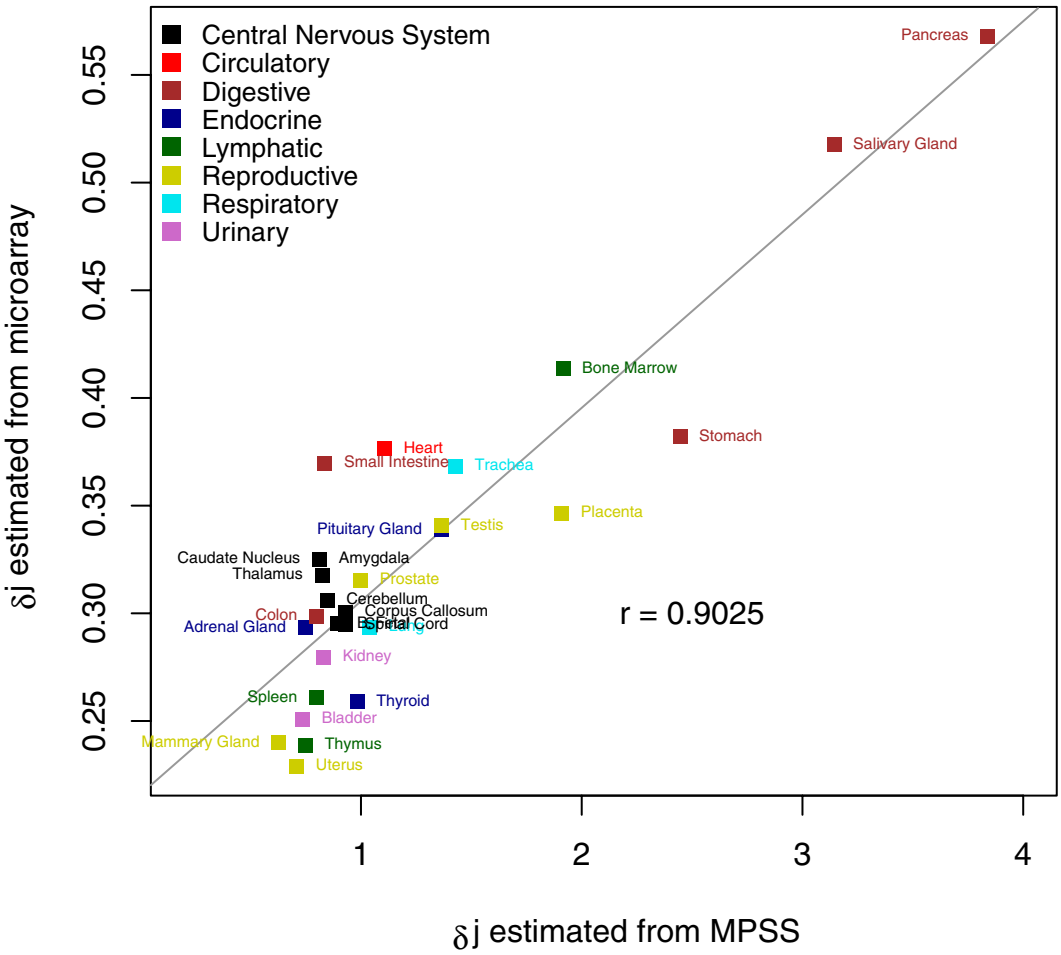

Fig. S2. Scatter plot for the estimated values of δ_j from the MPSS (*x* axis) and microarray dataset (*y* axis) in the 28 shared tissues between datasets. Tissues are colored by system of origin. The gray line indicates a lineal model fitted between the two variables. The value of the r Pearson's correlation coefficient is presented as text.

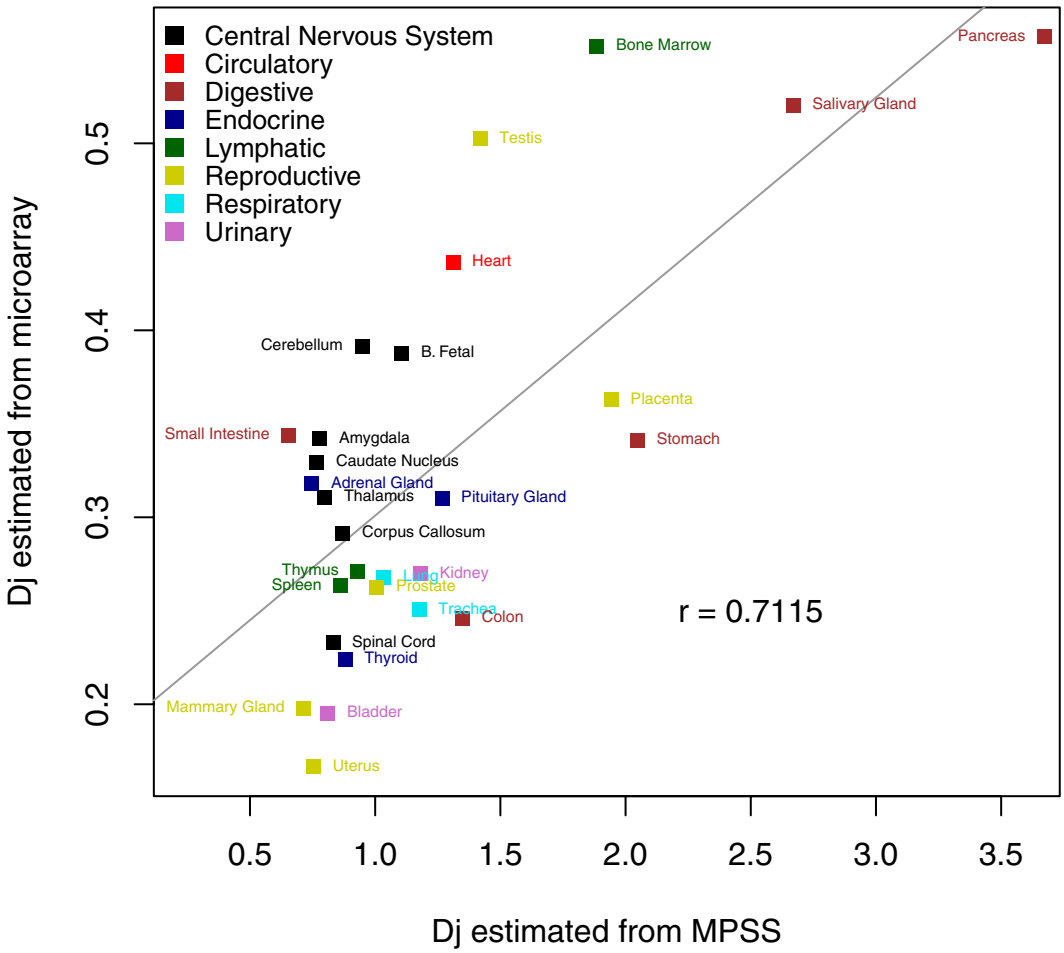

Fig. S3. Scatter plot for the estimated values of D_j from the MPSS (x axis) and microarray dataset (y axis) in the 28 shared tissues between datasets. Tissues are colored by system of origin. The gray line indicates a lineal model fitted between the two variables. The value of the r Pearson's correlation coefficient is presented as text.

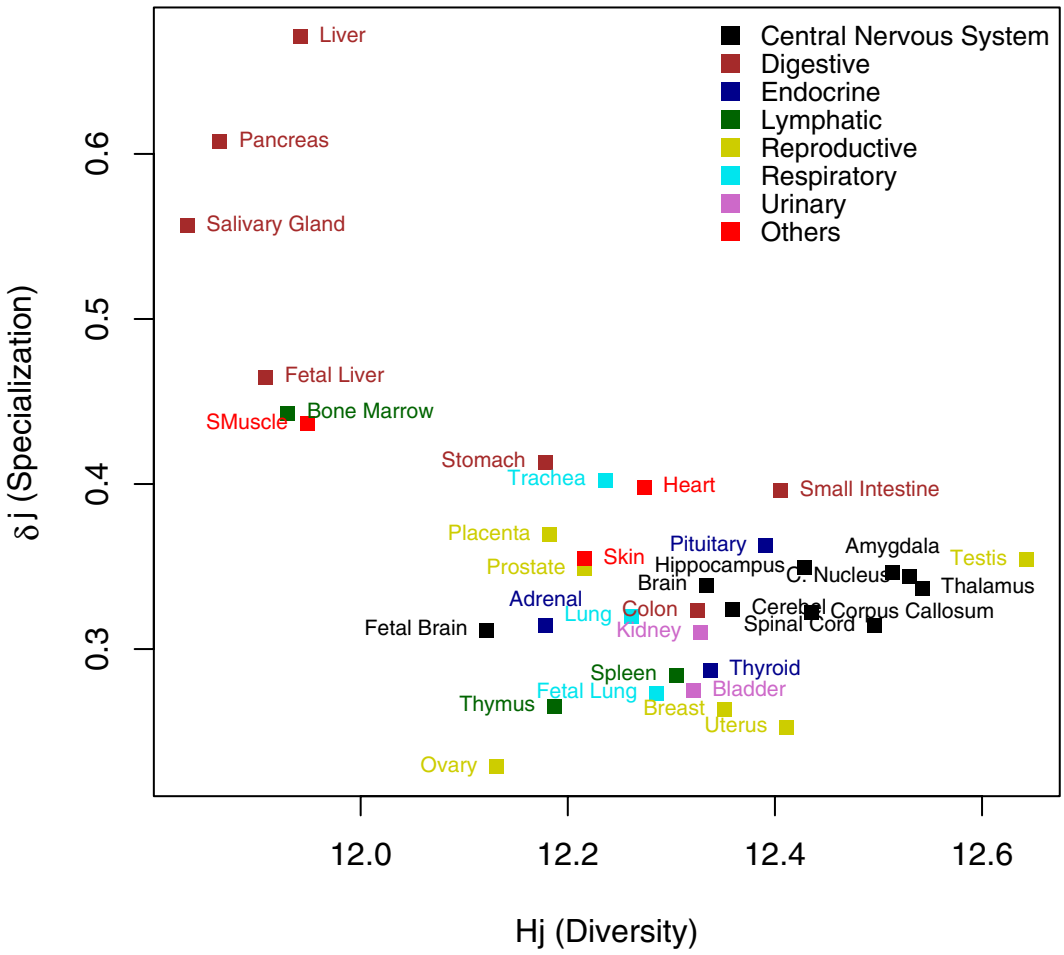

Fig. S4. Scatter plot of estimated values of H_j (diversity) vs. δ_j (specialization, given by the average gene specificity) for the 36 tissues of the human systems included in the microarray dataset. Tissues are colored by system of origin.

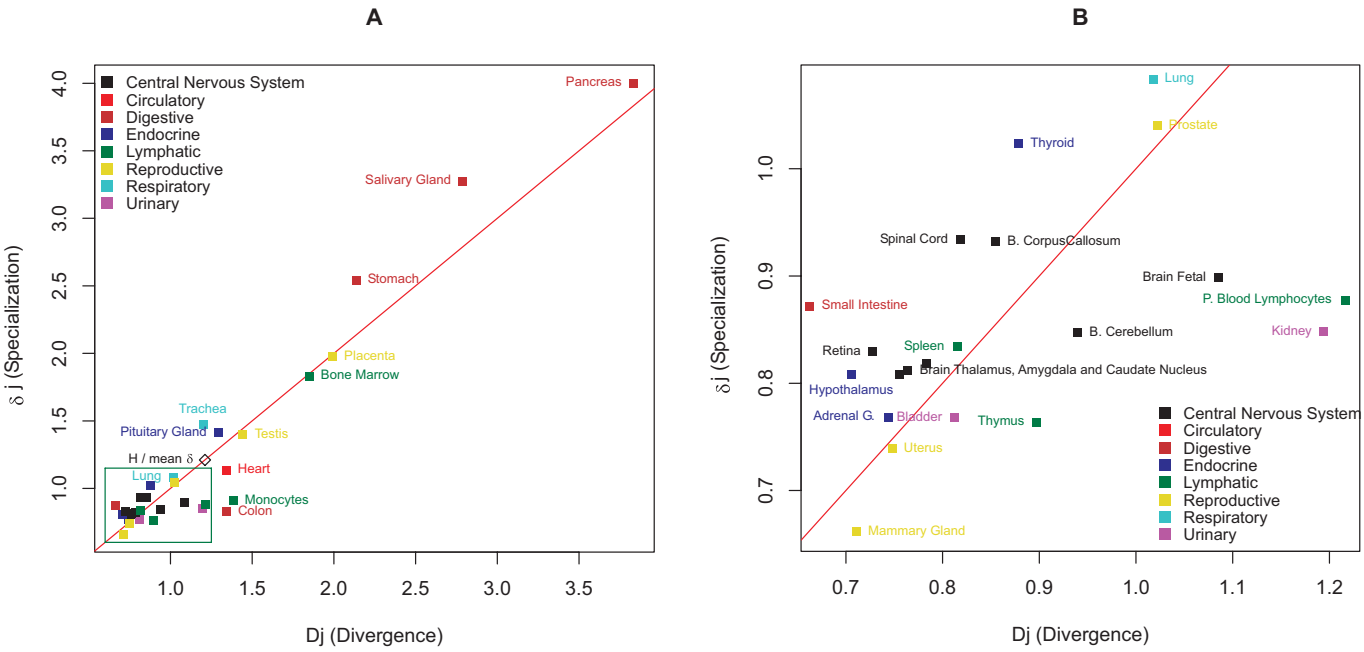

Fig. S5. Scatter plot of estimated values of D_j (divergence) *versus* δ_j (specialization) for tissues of the human systems obtained from the MPSS dataset. Tissues are colored by system of origin. A red line indicates the values where $D_j = \delta_j$. *B* is an amplification of the box in *A*.

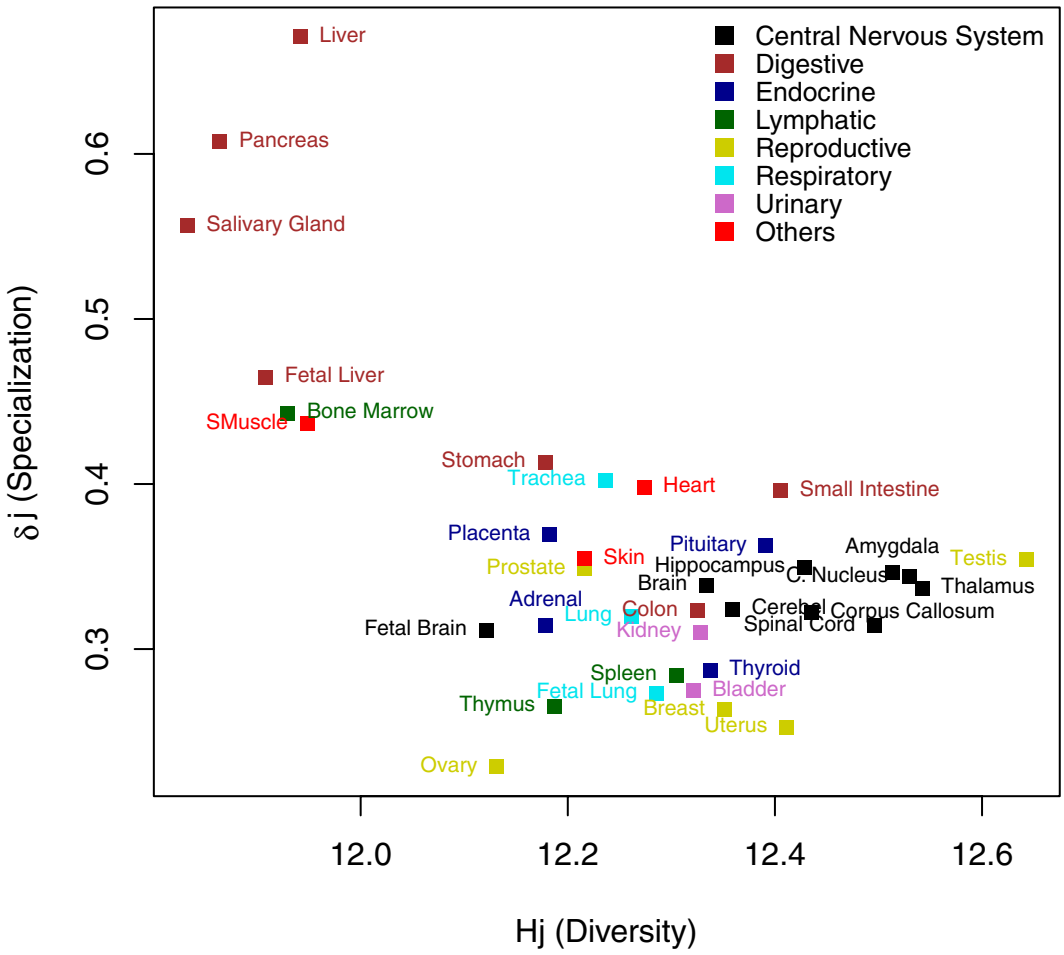

Fig. S6. Scatter plot of estimated values of H_j (diversity) *versus* δ_j (specialization, given by the average gene specificity) for the 36 tissues of the human systems included in the microarray dataset. Tissues are colored by system of origin.

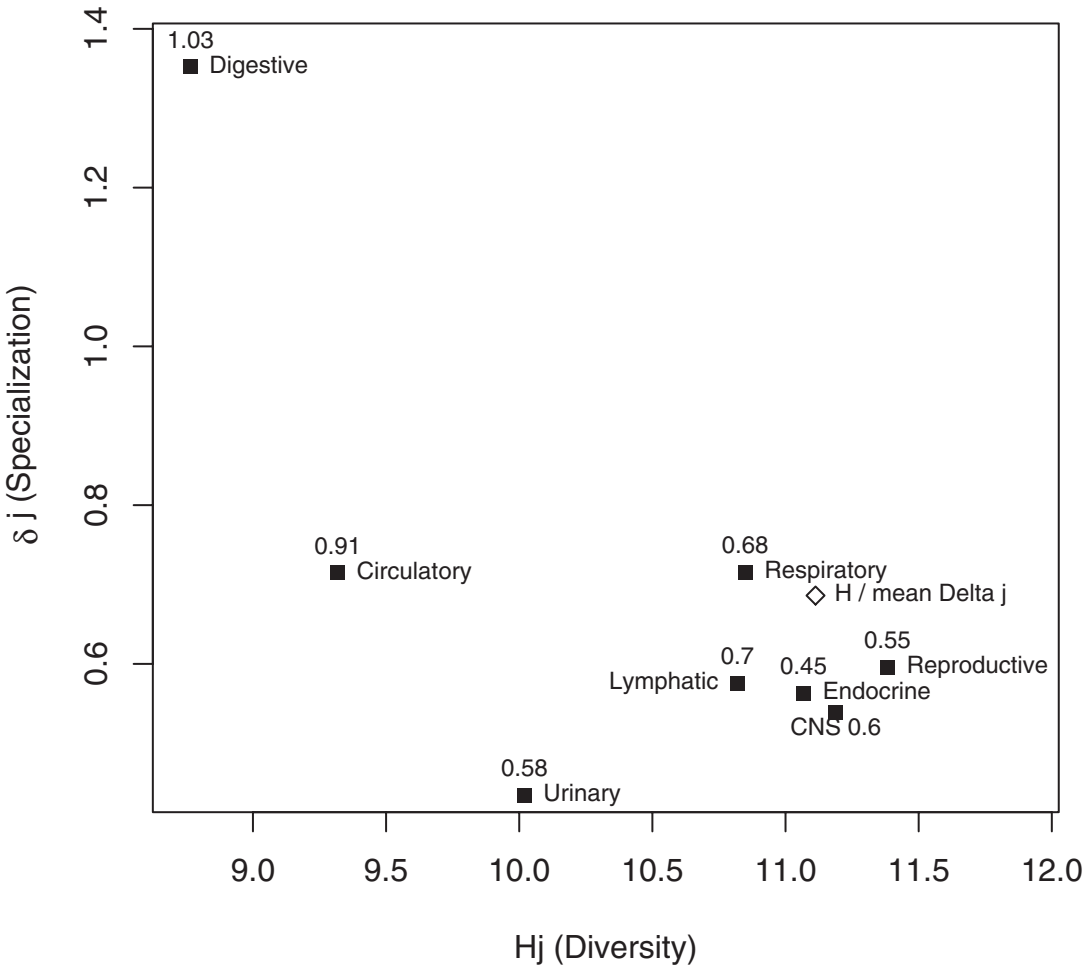

Fig. S7. Scatter plot of estimated values of H_j (diversity) vs.δ_j (Specialization, given by the average gene specificity) for each human system using the MPSS dataset. The values of D_j (divergence) for each tissue are shown close to each data point. The point of H for the whole system and the mean of δ_j is shown by a diamond.

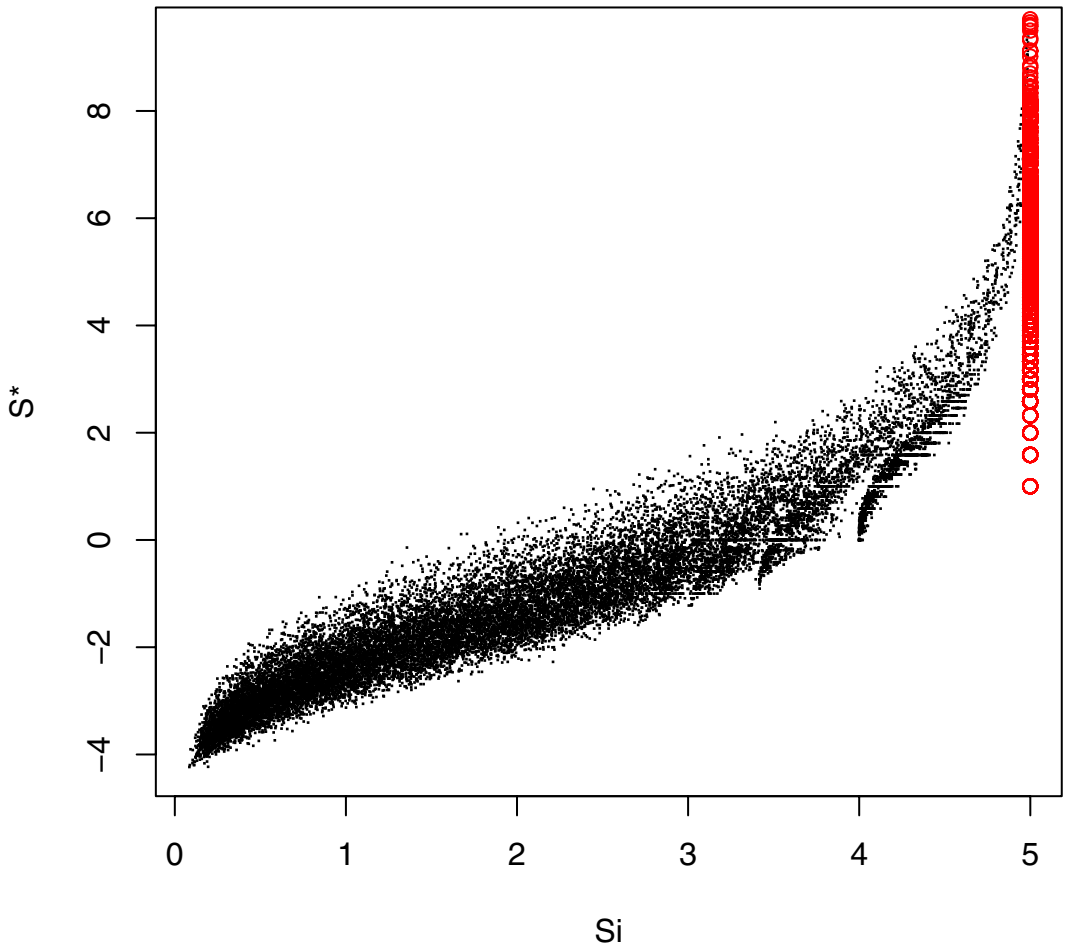

Fig. S8. Scatter plot for the values of the specificity coefficient proposed in Jongeneel *et al.* [Jongeneel CV *et al.* (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* 15:1007–1014], S*, and our specificity coefficient S_i for the genes in the MPSS human dataset. Red symbols are used for genes that reached the maximum value of S_i. Notice how the rank in S* of the specific genes ($S_i = 5$) includes genes with low values for S*, implying misclassification of specific genes when using the S* coefficient.

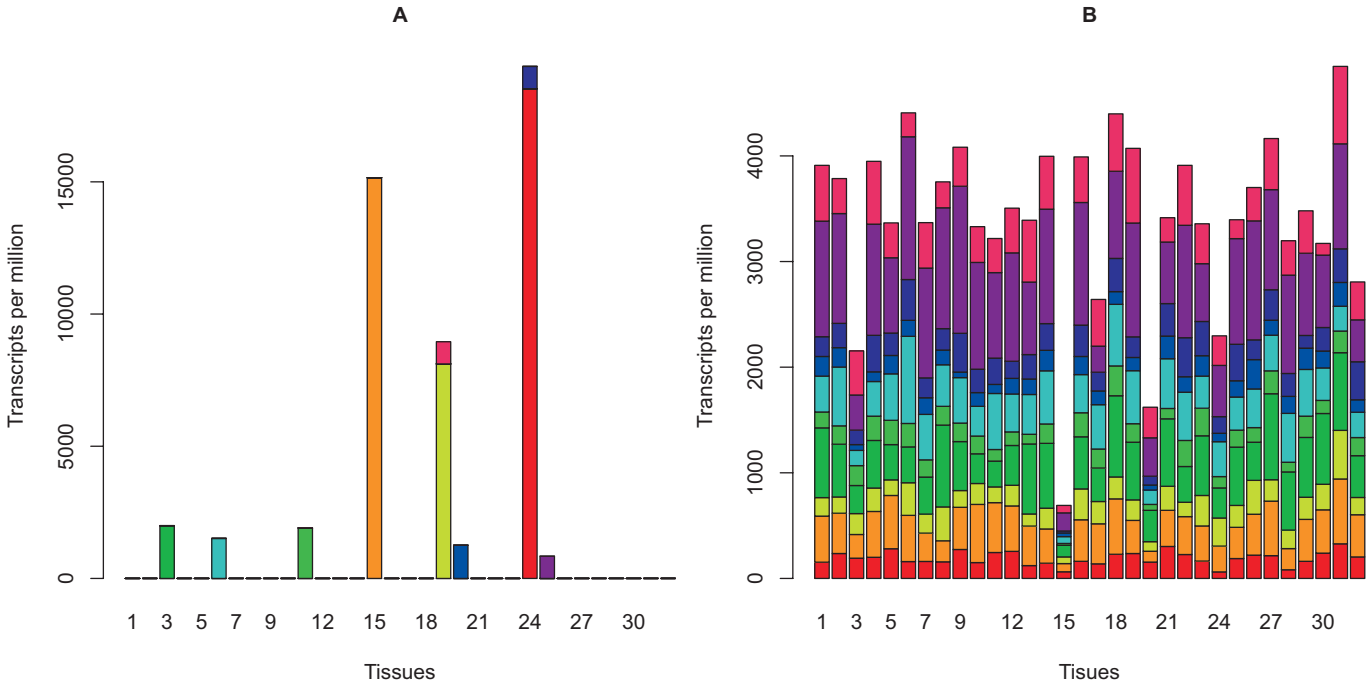

Fig. S9. Bar diagrams for the relative frequency of transcription in transcripts per million of 10 genes with the maximum values of S_i ($S_i = 5$; *A*) and the 10 genes with the lowest values of S_i (*B*). Data resulted from the analysis of the MPSS dataset. Tissues are numbered from 1 to 32 in the *x* axis. Relative frequencies of each gene are represented in a distinct color for each gene in each image independently.

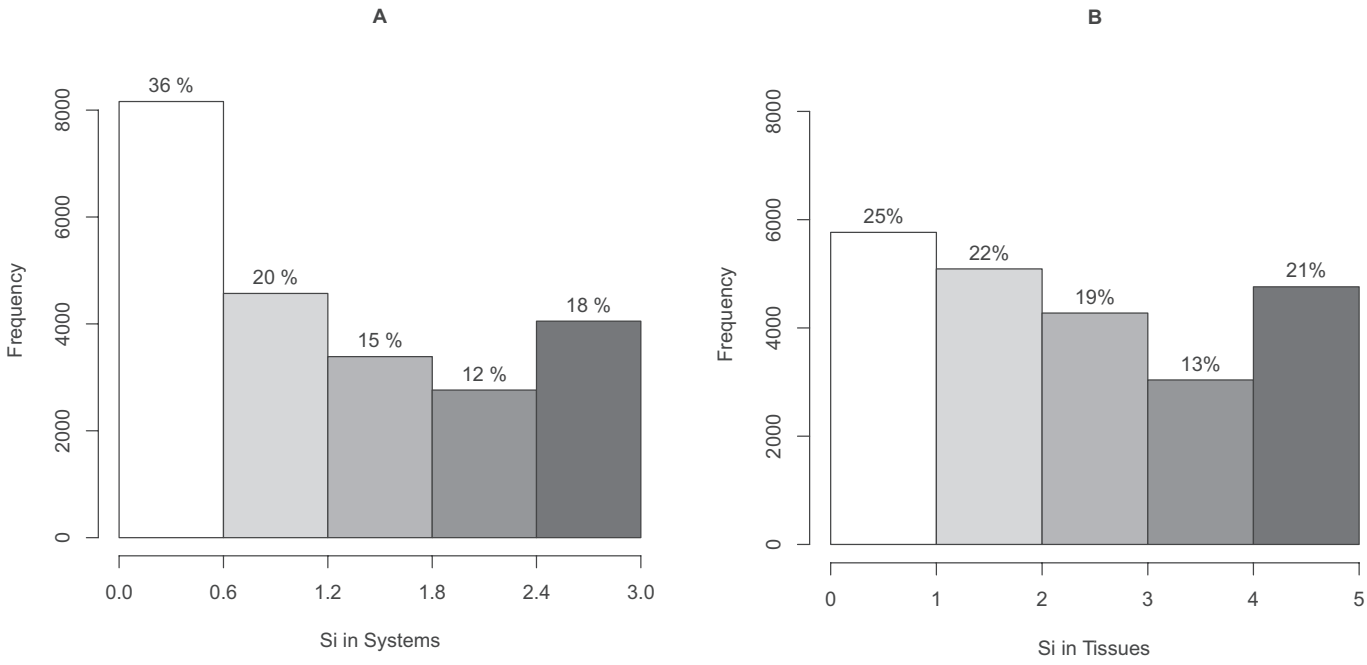

Fig. S10. Histograms for the values of gene specificity, S_i, when the tissues are grouped into systems (*A*) or are taken at tissue level (*B*) in the MPSS dataset. To make the graphs comparable the same number of classes is used in both cases and the frequency scale is the same. Percentages of observatons grouped in each class are shown above every bar. Gray tones are used from left to right to remark higher values of specificity.

Table S1. Information properties of human systems and tissues estimated from the MPSS dataset

System/tissue	Tags	*Hj*	*δj*	*Dj*
CNS	**9,818,779**	**11.1892**	**0.5385**	**0.5982**
Brain amygdala	973,569	10.8043	0.8115	0.7641
Brain caudate nucleus	1,039,527	10.2568	0.8079	0.7550
Brain cerebellum	902,231	10.8696	0.8471	0.9398
Brain corpus callosum	882,445	10.8472	0.9322	0.8547
Brain fetal	767,919	11.5496	0.8982	1.0850
Brain thalamus	1,076,935	10.2647	0.8184	0.7836
Retina	929,482	11.3452	0.8295	0.7275
Spinal cord	1,046,877	10.4470	0.9340	0.8187
Circulatory	**1,099,897**	**9.3186**	**0.7147**	**0.9104**
Heart	1,099,897	9.3186	1.1339	1.3404
Digestive	**5,145,463**	**8.7652**	**1.3521**	**1.0310**
Colon	1,028,538	8.5206	0.8289	1.3411
Pancreas	1,122,006	5.1984	4.0019	3.8331
Salivary gland	1,010,284	6.3352	3.2754	2.7890
Small intestine	995,657	10.6991	0.8718	0.6627
Stomach	988,978	6.9693	2.5434	2.1350
Endocrine	**4,074,576**	**11.0680**	**0.5629**	**0.4455**
Adrenal gland	1,054,497	10.4812	0.7683	0.7438
Brain hypothalamus	983,980	10.7645	0.8084	0.7056
Pituitary gland	1,068,227	10.1772	1.4137	1.2957
Thyroid	967,872	10.8784	1.0235	0.8786
Lymphatic	**4,553,040**	**10.8201**	**0.5752**	**0.7005**
Bone marrow	1,010,511	8.1848	1.8321	1.8479
Monocytes	875,905	10.7058	0.9081	1.3827
Peripheral blood lymphocytes	879,033	9.4792	0.8776	1.2167
Spleen	923,030	11.2090	0.8342	0.8149
Thymus	864,561	11.5884	0.7635	0.8969
Reproductive	**4,887,892**	**11.3831**	**0.5955**	**0.5489**
Mammary gland	1,037,119	10.6225	0.6615	0.7108
Placenta	939,372	9.0287	1.9777	1.9911
Prostate	1,012,568	10.7959	1.0404	1.0219
Testis	956,455	11.7426	1.3973	1.4422
Uterus	942,378	11.3665	0.7388	0.7478
Respiratory	**2,010,992**	**10.8485**	**0.7153**	**0.6772**
Lung	994,976	10.9640	1.0836	1.0177
Trachea	1,016,016	10.1933	1.4710	1.1992
Urinary	**2,021,104**	**10.0194**	**0.4348**	**0.5773**
Bladder	978,049	10.3861	0.7678	0.8119
Kidney	1,043,055	9.0806	0.8479	1.1942

Tags, total number of gene tags sampled in the system or tissue. H_j, *measure of diversity;* δ_j, *average gene specificity;* D_j, divergence with the average transcriptome.