Briefings in Bioinformatics, 18(2), 2017, 205–214

doi: 10.1093/bib/bbw008
Advance Access Publication Date: 18 February 2016
Paper

A benchmark of gene expression tissue-specificity metrics

Nadezda Kryuchkova-Mostacci and Marc Robinson-Rechavi

Corresponding author. Marc Robinson-Rechavi, Department of Ecology and Evolution, Batiment Biophorem, UNIL-Sorge, 1015 Lausanne, Switzerland. Tel.: +41-21-692-4220; E-mail: marc.robinson-rechavi@unil.ch

Abstract

One of the major properties of genes is their expression pattern. Notably, genes are often classified as tissue specific or housekeeping. This property is of interest to molecular evolution as an explanatory factor of, e.g. evolutionary rate, as well as a functional feature which may in itself evolve. While many different methods of measuring tissue specificity have been proposed and used for such studies, there has been no comparison or benchmarking of these methods to our knowledge, and little justification of their use. In this study, we compare nine measures of tissue specificity. Most methods were established for ESTs and microarrays, and several were later adapted to RNA-seq. We analyse their capacity to distinguish gene categories, their robustness to the choice and number of tissues used and their capture of evolutionary conservation signal.

Key words: tissue specificity; expression; human; mouse; RNA-seq; microarray

Introduction

Gene expression analysis is widely used in genomics and measured with microarrays or RNA-seq. In the case of a multicellular organism with different tissues, it is often useful to have a measure of how tissue specific a gene is.

Even if tissue specificity is often used in studies, there is usually no clear answer why one or another method was used. Yet, there are several methods to measure gene specificity, which differ in their assumptions and their scale. The simplest one is to count in how many tissues each gene is expressed (used in e.g. [1–6]). The problem of this method is to define the threshold to call a gene expressed. Originally, with expressed sequence tags (ESTs), a count of 1 EST was considered sufficient [2]. There are different methods to define thresholds for microarrays [7], while for RNA-seq, an Reads Per Kilobase per Million mapped reads (RPKM) value of 1 is generally used [8, 9]. Some studies use a stringent threshold, e.g. signal to noise ratio >10 [10], and count a gene as specific only if expressed in a single tissue. This method causes only highly expressed genes to be taken into account, and if a data set contains closely related tissues (e.g. brain parts), less genes are called tissue specific. Other papers a very low threshold, e.g. 0.3 RPKM [11–13], that leads to defining most genes as housekeeping.

A widely used method which does not depend on such a cut-off in its formula is Tau [14] (for details, see 'Materials and Methods'). Tau varies from 0 to 1, where 0 means broadly expressed, and 1 is specific (used in e.g. [15–24]).

Other methods have been proposed, such as the expression enrichment (EE) [25], to calculate for which tissue each gene is specific, for example, in the database TiGER [26]. We also considered: the tissue specificity index (TSI) [27] (used in e.g. [28–30]); Hg by Schug *et al.* [31]; the z-score (used in [32]), widely used for other features than tissue specificity; SPM, used in the database TiSGeD [33]; and Preferential Expression Measure (PEM), suggested for ESTs by Huminiecki *et al.* [34] and used in e.g. [35–38]. Finally, the Gini coefficient, widely used in economics to measure inequality [39], was compared with methods originating in biology.

These methods can be divided in two groups. One group summarizes in a single number whether a gene is tissue specific or ubiquitously expressed (Tau, Gini, TSI, Counts and Hg), and

Nadezda Kryuchkova-Mostacci is a doctoral student in biology in the Department of Ecology and Evolution, University of Lausanne. Her main interest is in understanding the relation between gene expression patterns and molecular evolution.
Marc Robinson-Rechavi is an associate professor at the Department of Ecology and Evolution in the University of Lausanne, and Group Leader at the Swiss Institute of Bioinformatics. His main interest is in the evolution of animal genomes in the context of organismal function and development.
Submitted: 11 December 2015; **Received (in revised form):** 6 January 2016

the second group shows for each tissue separately how specific the gene is to that tissue (z-score, SPM, EE and PEM). For comparison purposes with the first group, we use the maximum specificity from the second group.

Material and methods

For all equations,

x_i *is the expression of the gene in tissue* i

n *is the number of tissues*

The method of counting in how many tissues a gene is expressed was simply calculated as follows:

$$Counts = \sharp \text{ tissues expressed}$$

A cut-off needs to be set; the cut-offs that we used are explained at the end of the 'Methods' section.

While the other methods do not necessitate a cut-off per their mathematical formulation, they need positive expression values. As expression values are usually log-transformed (because they are log-normally distributed), this means that values <1 are not manageable. Solutions include using a multiplier or/and setting a cut-off of 1 before log transformation. For details of our treatment of the data, see the description of RNA-seq and microarray data.

Tau was calculated as follows [14]:

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x_i})}{n-1};\ \hat{x_i} = \frac{x_i}{\max\limits_{1\le i\le n}(x_i)}.$$

The EE score was calculated as follows [25]:

$$EE = \frac{x_i}{\sum_{i=1}^{n} x_i * \frac{s_i}{\sum_{i=1}^{n} s_i}} = \frac{\sum_{i=1}^{n} s_i}{s_i} * \frac{x_i}{\sum_{i=1}^{n} x_i}$$

s_i *summary of the expression of all genes in tissue* i.

TSI was calculated as follows [27]:

$$TSI = \frac{\max\limits_{1\le i\le n}(x_i)}{\sum_{i=1}^{n} x_i}.$$

The Gini coefficient was calculated as follows:

$$Gini = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n}(n+1-i)x_i}{n\sum_{i=1}^{n} x_i}$$

x_i *has to be ordered from least to greatest.*

Hg [31] was calculated as follows:

$$H_g = -\sum_{i=1}^{n} p_i * \log_2(p_i);\ p_i = \frac{x_i}{\sum_{i=1}^{n} x_i}.$$

The z-score was calculated as follows:

$$z = \frac{x_i - \mu}{\sigma}$$

μ *is the mean of gene expression;* σ *is the standard deviation.*

The z-score can be implemented in two ways: either only over-expressed genes are defined as tissue specific, or the absolute distance from the mean is used, so that under-expressed genes are also defined as tissue specific. Only the former method was used to be able to compare z-score with other methods.

SPM from the database TiSGeD [33] was calculated as follows:

$$SPM = \frac{x_i^2}{\sum_{i=1}^{n} x_i^2}.$$

PEM estimates how different the expression of the gene is relative to an expected expression, under the assumption of uniform expression in all tissues. PEM is calculated as follows [34]:

$$PEM = \log_{10}\left(\frac{\sum_{i=1}^{n} s_i}{s_i} * \frac{x_i}{\sum_{i=1}^{n} x_i}\right)$$

s_i *summary of the expression of all genes in tissue* i.

Derivation of all the methods from the original equations is presented in Supplementary Materials.

The output of all methods was modified to the same scale from 0 (ubiquitous) to 1 (tissue specific) to be able to compare them (Table 1). Four of the methods calculate specificity value for each tissue separately; for these methods, the largest (most specific) value among all tissues was assigned to the gene (see Table 1).

All the methods were compared using R version 3.2.1 [40], with the gplots [41], reldist [42, 43], VennDiagram [44] and preprocessCore libraries [45]; the R script is available in Supplementary Materials.

We used the following RNA-seq data: 27 human tissues (E-MTAB-1733) from Fagerberg *et al.* [46] downloaded from their Supplementary Materials, 22 mouse tissues (GSE36025) from the ENCODE project [47, 48] as used in Kryuchkova-Mostacci and Robinson-Rechavi [20] and 8 human tissues and 6 mouse tissues from Brawand *et al.* [49], as processed in the Bgee database [50]. All the genes with expression <1 RPKM were set as not expressed. The RNA-seq data were first log-transformed. After the normalization, a mean value from all replicates for each tissue separately was calculated. All genes that were not expressed in at least one tissue were removed from the analysis.

We used the following microarray data, as annotated in the Bgee database: 32 human tissues (GSE2361) [51] and 19 mouse tissues (GSE9954) [52]. Of note, on the microarrays, we have only 9788 (resp. 16 043) genes with data in human (resp. mouse), relative to 18 754 (resp. 27 364) for RNA-seq. For the microarray data, we used the logarithm of normalized signal intensity. The values set as absent in Bgee were set to 0, following the method of Schuster *et al.* [53]. After the normalization, a mean value from all replicates for each tissue separately was calculated. All genes that were not expressed in at least one tissue were removed from the analysis.

A summary of the workflow is presented in Supplementary Figure S1.

For the comparison of tissue-specific or ubiquitous gene functions, we used the following Gene Ontology (GO) terms: spermatogenesis (GO:0007283; expected to be specific to testis; 469 human genes), neurological system process (GO:0050877; expected to be specific to brain and other neural tissues; 1338 human genes), xenobiotic metabolic process (GO:0006805;

Table 1. Tissue specificity parameters. *N* is the number of tissues in the data set

Methods	Tissues	Ubiquitous	Specific	Transformation
τ (*tau*)	all	0	1	–
Gini	all	0	$(N-1)/N$	$x*(N/(N-1))$
TSI	all	0	1	–
Counts	all	N	1	$(1-x/N)*(N/(N-1))$
EE_i	separately	0	>5	X/maxX
Hg	all	$\log_2 N$	0	$1-x/\log_2 N$
Z score	separately	0	>3	$X/n-1/\sqrt{N}$
PEM score	separately	0	~ 1	X/max X
SPM	separately	0	1	X

$X = \max_{1\le i\le n} x_i$ is the maximal specificity value for a certain gene among all tissues.

expected to be specific to liver and kidney; 163 human genes), protein folding (GO:0006457; expected to be ubiquitous; 231 human genes), membrane organization (GO:0061024; expected to be ubiquitous; 607 human genes) and RNA splicing (GO:0008380; expected to be ubiquitous; 383 human genes).

GO enrichment analysis was performed with GOrilla [54] and Revigo [55].

Results

All methods show a bimodal distribution of gene scores: most genes are either broadly expressed or specific, with only few in between. This is true both with RNA-seq data (Figure 1 and Supplementary Figure S2) and with microarray data (Supplementary Figures S3 and S4). Most methods are strongly skewed towards classifying many genes as ubiquitous, and few as tissue specific or intermediate. Z-score has a shifted peak of tissue specificity relative to other metrics. Tau has a less skewed distribution, with the most tissue-specific and intermediate genes, indicating that it might be capturing more of the variance among gene expression patterns.

All methods correlate relatively well with each other (Supplementary Figures S5 and S6), but the relation is often not linear because methods other than Tau and Gini have little variance outside of the most tissue-specific genes. For example, genes which have Tau between 0.85 and 0.95 have Tsi between 0.2 and 0.43 (Supplementary Figure S5).

As a first measure of robustness of tissue-specificity metrics, we compared each metric calculated on the full human RNA-seq data set of 27 tissues, and on subsets of five tissues (Figure 2). Not all permutations were performed, for computational reasons, but a random sample of 1000 permutations. Ideally, the signal for tissue specificity should already be detectable with the five tissues. Tau, Gini, Counts, PEM and the Hg coefficient all show correlations which are not too low (mean $r > 0.4$), indicating that these methods are reasonably robust to the number of tissues. TSI, SPM and EE score show weaker results (mean $0.2 > r > 0$). The correlation for z-score is even negative, indicating that it should be not used with a small number of tissues, and casting doubt on its utility to robustly estimate tissue specificity. We performed the same analysis in mouse, comparing scores between all 22 available tissues and subsets of five tissues; the results are consistent, but correlations are weaker for all parameters (Supplementary Figure S7). Similarly, we compared the scores using all available tissues (27 in human, 22 in mouse) with the scores using only the 16 tissues shared between these human and mouse data sets; correlations of all parameters are high for human and mouse, and z-score shows again the lowest correlation in all cases (Supplementary Figures S8 and S9).

The choice of tissues to calculate tissue specificity affects the results. All the outliers (stronger correlation) in Figure 2 and Supplementary Figure S7 contain testis tissue. This can be explained by the fact that testis has the largest number of tissue-specific genes (Supplementary Figures S10 and S11). Thus, using a subset which excludes testis produces an estimate of tissue specificity that is biased relative to the full data set, and this bias is only relieved in the few subsets that include testis.

We also analysed robustness of Tau by comparing correlation calculated on all 27 tissues and on all the subsets of 5–26 tissues (Supplementary Figures S12 and S13). Again, all the subsets that are most similar to the full set (outliers with $r > 0.7$) in subsets of five and six tissues contain testis in the set. Conversely, all the subsets that are most different in the full set (outliers with $r < 0.8$) in subsets of 21–26 tissues do not have testis in the subset. There are other outlier subsets that are closer to the main distribution for 25 or 26 tissues: these do include testis, but not brain, which is the second tissue with the most specific genes.

In addition to being robust to tissue sampling, we expect a good measure of tissue specificity to capture biological signal. A simple expectation of such biological signal is that it should be mostly conserved between orthologues from closely related species such as human and mouse [56]. Thus, we compared the methods in their conservation between human and mouse, using the 16 common tissues (Figure 3). All of the methods, except z-score, show a high correlation ($r > 0.69$). Specificity parameters calculated on only six common tissues between mouse and human (from the Brawand *et al*. data set) show even higher correlations ($r > 0.75$, Supplementary Figure S14).

Another way to capture biological signal is to compare the expression specificity of genes annotated with functions that are expected to be tissue specific, or which are expected to be ubiquitous. For this, we chose three tissue-specific GO terms and three GO terms that are expected to be present in all tissues. The tissue-specific GO terms are spermatogenesis, specific to testis; neurological system process, specific to brain and other neural tissues; and xenobiotic metabolic process, specific to liver and kidney. The broadly expressed GO terms are protein folding, membrane organization and RNA splicing. The distribution of the genes belonging to each category is presented in Figure 4 and Supplementary Figure S15. All of the parameters are successful at recognizing broadly expressed genes (peak of blue lines, as expected, shifted towards 0). But, there are important differences in results for specific genes. Only Tau has a

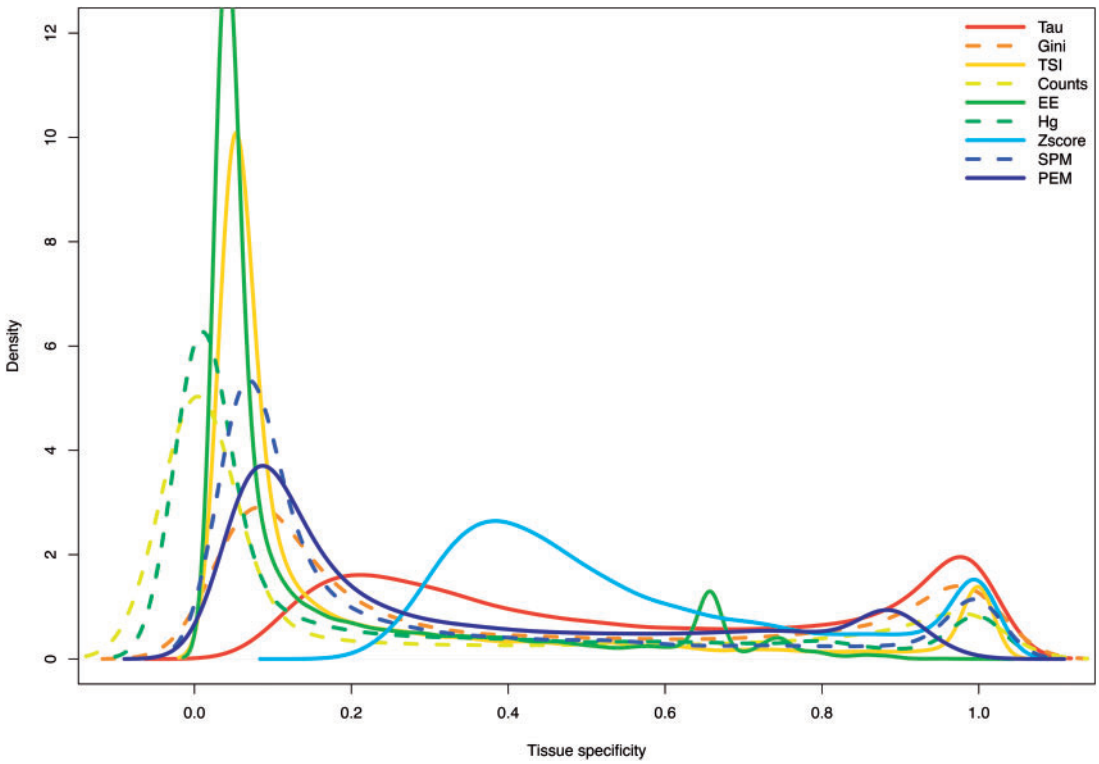

Figure 1. Distribution of tissue-specificity parameters with data for human RNA-seq of 27 tissues. Graph created with density function from R, which computes kernel density estimates. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

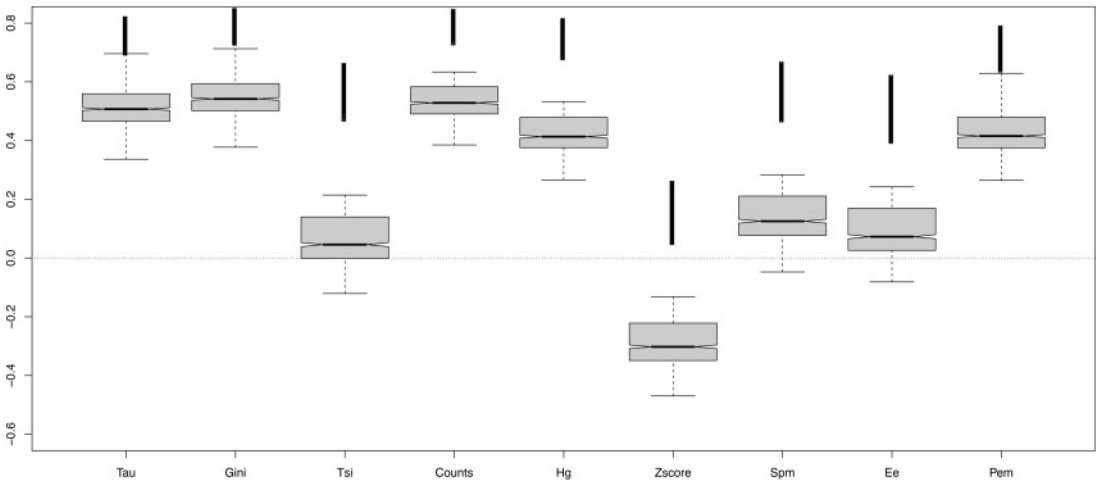

Figure 2. Comparison between tissue-specificity parameters calculated on the same human RNA-seq data set using all 27 tissues versus 1000 random subsets of five tissues.

larger peak close to 1 than close to 0. All parameters, except Tau, show strongly bimodal distributions for the genes that are expected to be specific, often with the larger peak at ubiquitous expression. Thus, Tau appears to be more successful at recovering this expected biological signal. We also checked the correlation of genes from tissue-specific functions (according to the three GO terms) between mouse and human orthologues (Supplementary Figure S16). Even if correlations are high and almost the same for all the parameters, the difference is that genes that are expected to be specific are specified by most parameters as ubiquitous. Only Tau reports most of these genes as evolutionarily conserved tissue specific.

Most methods seem to have more difficulty in finding tissue-specificity signal than broad expression signal. We checked whether those tissue-specific genes detected by each method are specific to the method, or also detected by others. Strikingly, almost all tissue-specific genes found by any method are also found by Tau. Gini also reports many tissue-specific genes that are reported by Tau but no other method. This is illustrated with the examples of brain- and testis-specific genes in Figure 5 (for other organs, see Supplementary Figure S17–S41). To call genes specific, a threshold of 0.8 was set, which is after the first peak of the bimodal distribution for most parameters. The same analysis was performed with thresholds of 0.6 and 0.4 (data not

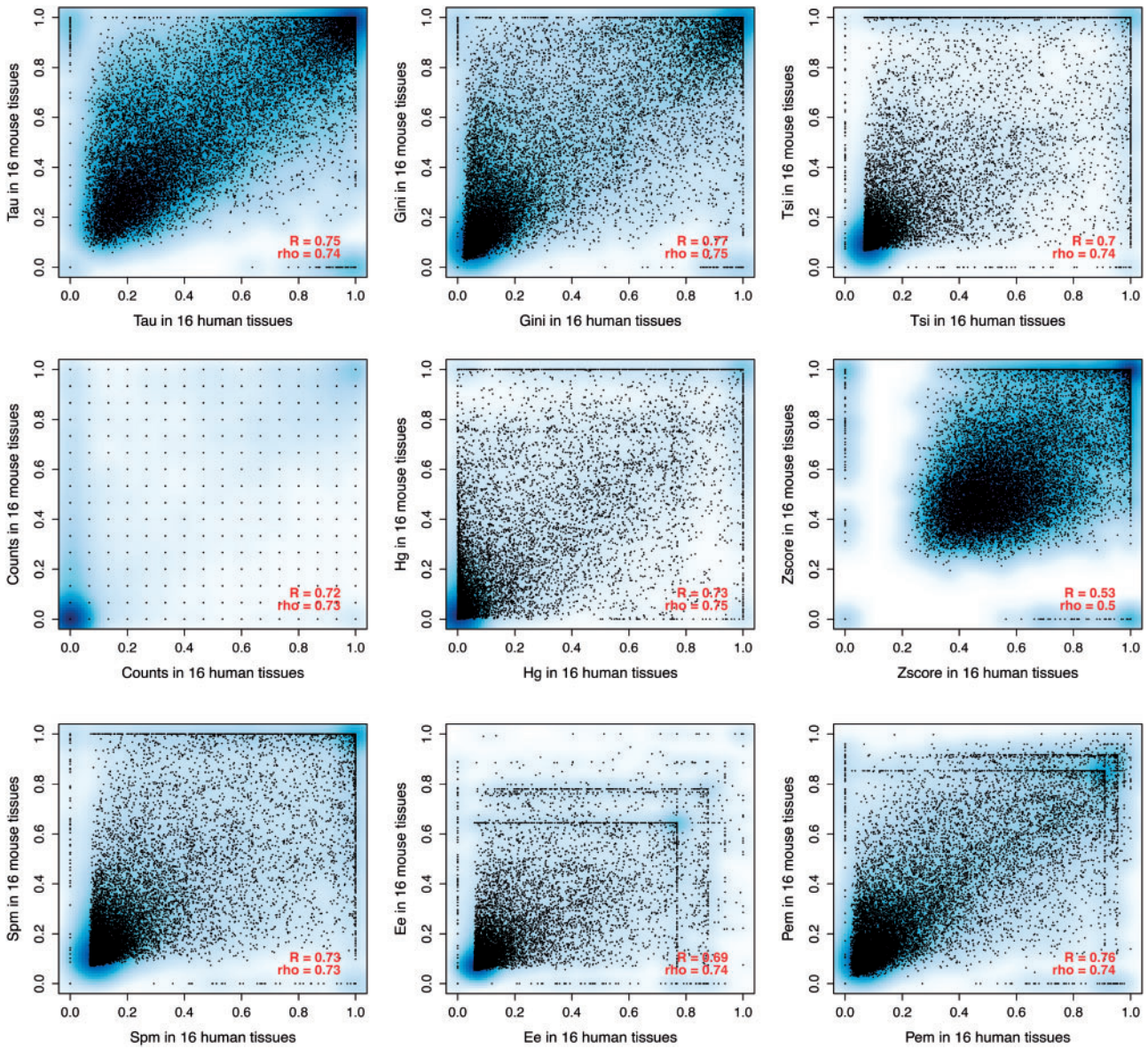

Figure 3. Comparison between tissue-specificity parameters calculated on the 16 common tissues between the human and mouse RNA-seq data sets. All correlations have P-value $<2.2 \times 10^{-16}$. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

shown), and produced similar results: Tau detects all genes that other methods detect plus some that are not detected by any other method. To check whether these additional tissue-specific genes found by Tau are biologically relevant, a GO enrichment test was performed on tissue-specific genes for testis and brain reported by all methods, by Tau alone or only by Tau and Gini (Supplementary Figures S42–S47). Each of these genes sets is indeed enriched in brain- or testis-specific functions, which shows that these were rather false negatives of the other methods than false positives of Tau and Gini.

The same analysis was also performed on the microarray data sets for mouse and human. We compared each metric on a full microarray human data set of 32 tissues and on the subset of five tissues (Supplementary Figure S48). For human, the correlations are weaker than with RNA-seq, even for the best performing metrics: Counts, Gini and Tau (mean $0.2 < r < 0.4$). For mouse, the correlations on microarray data are better: Counts, Gini and Tau (mean $0.4 < r < 0.6$) (Supplementary Figure S49). Results for 32 and 14 human tissues, and for 19 and 14 mouse tissues, are shown in Supplementary Figures S50 and S51. The distribution of correlations of Tau calculated on different subsets of tissues is shown in Supplementary Figures S52 and S53. Similarly, in the comparison between human and mouse orthologues, the correlations are much weaker for microarrays than for RNA-seq (Supplementary Figure S54). Specificity values are better correlated between RNA-seq and microarray for the mouse than for the human data sets (Figure 6 and Supplementary Figure S55). This correlation is on the same scale as that between two different RNA-seq data sets, although the correlation is a bit stronger for the RNA-seq data sets (Supplementary Figures S56 and S57). It should be noted that microarray and RNA-seq can only be compared on the subset of genes for which microarray data are usable, which excludes very tissue-specific genes detected only by RNA-seq (Supplementary Figure S58).

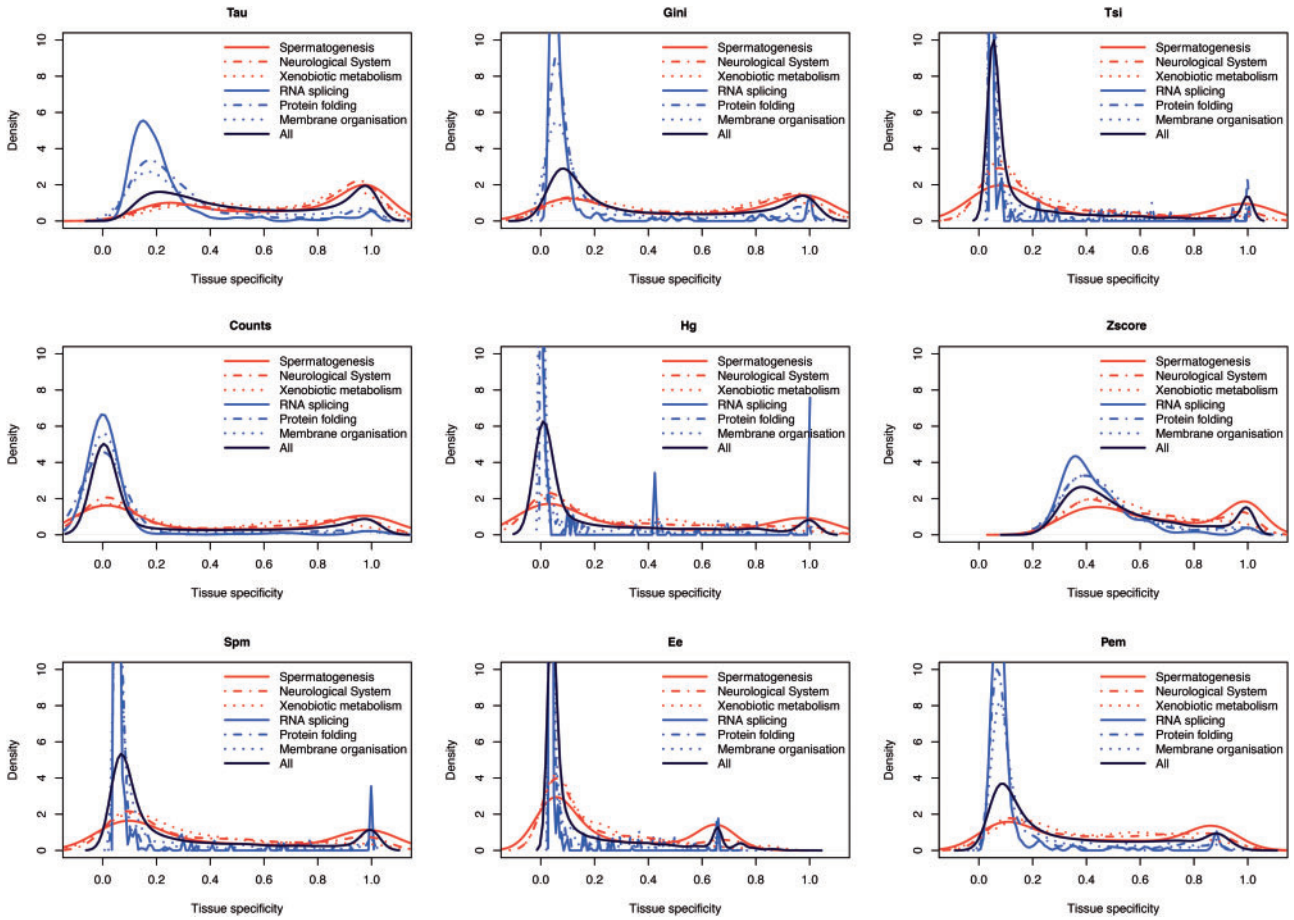

Figure 4. Tissue-specificity parameters of subsets of genes which are expected to be tissue-specific (top three terms, Spermatogenesis to Xenobiotic metabolism lines) or broadly expressed (RNA splicing to Membrane organisation lines), based on associated GO terms (described in Material and Methods). The black line represents the distribution for all genes, including those not associated to any of these GO terms. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Tissue specificity has been reported to be negatively correlated to mean or maximum gene expression level across tissues, i.e. ubiquitous genes have higher expression, and specific genes have lower expression (discussed in [1, 4, 20]). Indeed, we find a negative correlation of all metrics with mean expression; this correlation is similar for RNA-seq (r from -0.69 to -0.93) and for microarray (r from -0.70 to -0.95) (Supplementary Figures S59–S62). Z-score has the weakest correlation with mean expression on RNA-seq data and on microarray data. The correlation of tissue specificity parameters and maximal expression is also similar with RNA-seq and microarray (Supplementary Figures S63–S66): all the parameters are negatively correlated with maximal expression.

In all the analyses described above, RPKM values were log-transformed, as described in 'Material and Methods'. In the following, we investigated how stable the results of tissue specificity are if data are not log-normalized or if they are additionally quantile normalized. We compared tissue specificity calculated on log-transformed RPKM (as above), raw RPKM, log-transformed and quantile normalized RPKM (Supplementary Figures S67–S75); quantile normalization was performed across tissues in each data set. In general, quantile normalization has no influence on the results of calculation of tissue specificity (Supplementary Figures S74 and S75). Expectedly, removing log-transformation has a greater influence on all parameters, in the direction of detecting more tissue-specificity, sometimes losing completely the signal of broad expression, e.g. Tau (Supplementary Figure S68). Moreover, in the absence of log-transformation, the correlations between subsets of tissues or between species are in general weaker (Supplementary Figures S69–S70). The normalization has no influence on Counts, as expected, as only yes/no for the expression is taken in the account. Tau, Gini, TSI and Hg show the highest correlations between normalized and non-normalized data (Supplementary Figures S72 and S73), thus appearing more robust.

Discussion

We analysed nine parameters to calculate tissue specificity. We compared the methods with respect to their stability to the number of tissues, their correlation between one-to-one orthologues in human and mouse, their power in detecting tissue-specific genes and their distribution of values. As many experiments do not have many tissues, it is important that tissue specificity can be calculated reliably on few tissues.

Different methods of calculating tissue specificity take into account different properties of expression. The Counts method does not take in account the amplitude of differences between tissues. This is the simplest method; yet, if the threshold is chosen properly, it gives surprisingly good results. Distribution of Counts tissue specificity depending on the chosen threshold is presented in Supplementary Figure S76: with too high or too low threshold, most genes are reported as not specific, but it is robust to a change of one order of magnitude (1–10 RPKM). Tau and TSI

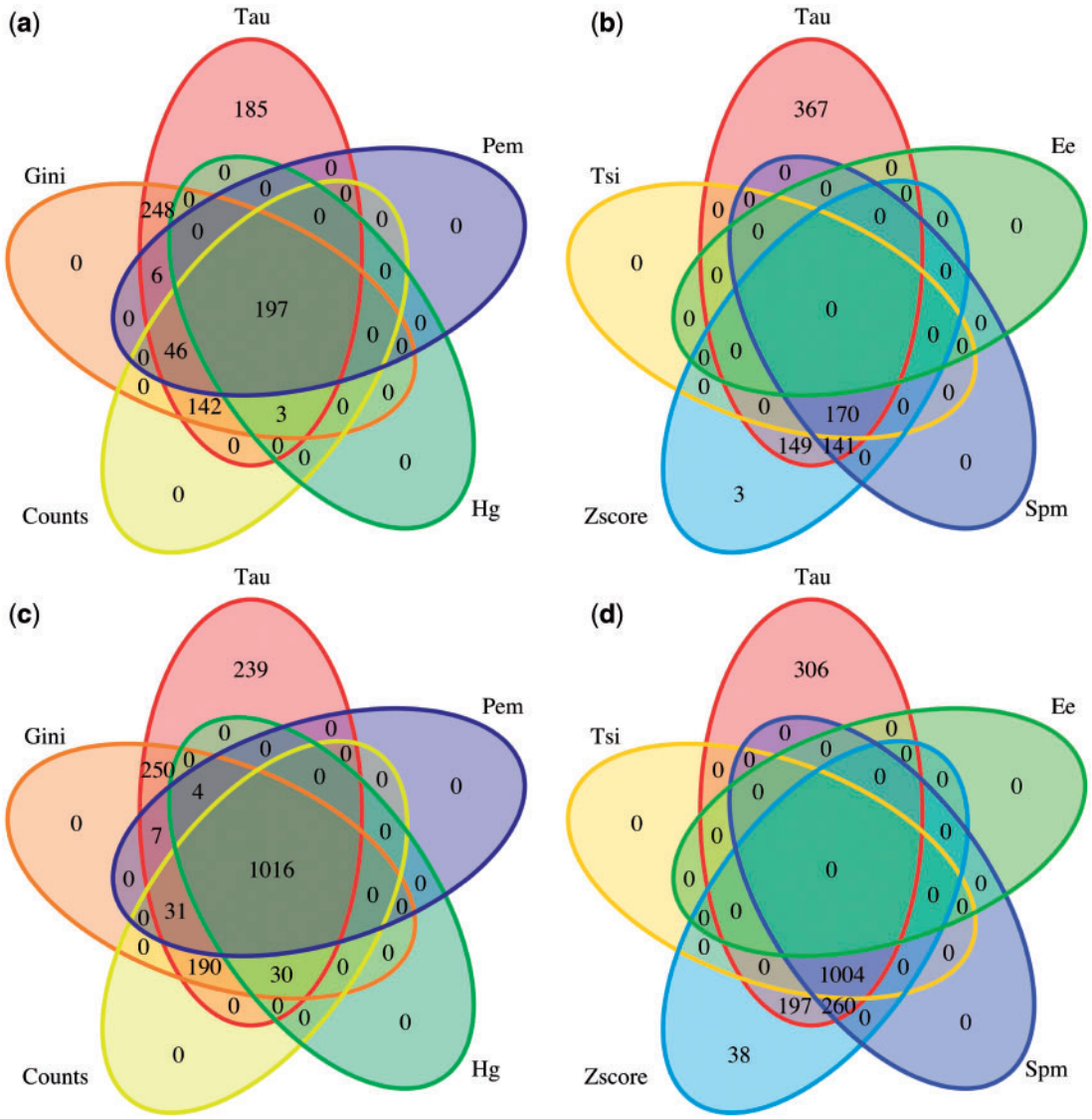

Figure 5. Venn diagram of genes called specific with different parameters, with a cut-off of 0.8 for each parameter; (**A**) and (**B**) genes with their highest expression in the brain; (**C**) and (**D**) genes with their highest expression in the testis; parameters are shown in A/C or B/D for readability, with Tau in common because it calls the most genes tissue specific. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

both use the information of expression of a gene in each tissue and its maximal expression over all tissues. The difference between Tau and TSI is that Tau also takes into account the number of tissues. The Hg coefficient is also similar, but differs in that instead of the maximal expression (necessarily in a specific tissue) the sum of expression over tissues is used, and each normalized value is multiplied by log of the value. And for the SPM score, each value (squared) is corrected by the sum of squared gene expression across all tissues. The EE score also corrects each expression value by the sum of gene expression across tissues as well as by the sum of expression in the target tissue. The PEM score is simply the logarithm (base 10) of the EE score. As these coefficients are normalized by either maximal expression of the gene or by the sum of expression of the gene, they are not sensitive to its absolute expression level. Z-score is the only method that takes the standard deviation of expression into account. An overview of the methods with their shared components (e.g. max expression appears in Tau and in TSI) is presented in the Supplementary Materials.

Tau appears consistently to be the most robust method in our analyses. Comparing coefficients calculated on different sized data sets, Tau showed one of the highest correlations (Figure 2 and Supplementary Figures S7–S9). And, while it may be debated what is it the 'best' distribution between ubiquitous and specific genes, we note that Tau provides well-separated groups with lower skew towards calling most genes ubiquitous or tissue specific than other methods (Figure 1 and Supplementary Figure S2); and it found more tissue-specific genes (Figure 5, Supplementary Figures S10, S11 and S16–S41). Tau also showed a robust behaviour according to normalization of data (Supplementary Figures S72 and S73). With the GO analysis performed, Tau is the best in recognizing tissue-specific genes (Figure 4 and Supplementary Figure S15), and conversely tissue-specific genes found only with Tau have functional annotations that are consistent with their tissue of highest expression (Supplementary Figures S41–44).

When a score per tissue is needed, the PEM score showed acceptable results, except for non-log-transformed mouse

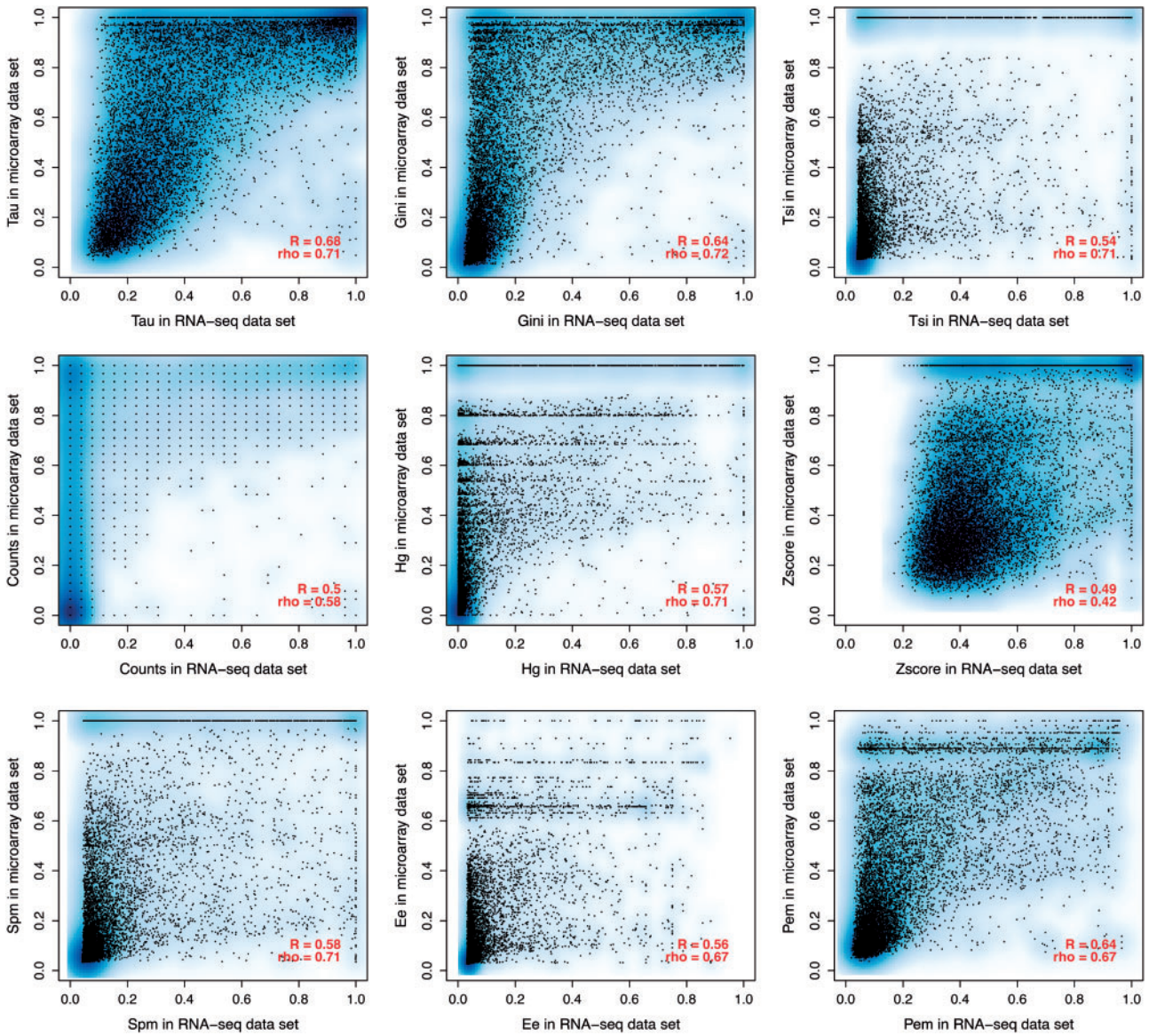

Figure 6. Comparison between tissue-specificity parameters calculated on RNA-seq of 27 tissues versus microarray of 32 tissue in human data sets. All correlations have *P*-value $<2.2 \times 10^{-16}$. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

RNA-seq (Supplementary Figure S73), and it is most similar to Tau. An association between scores and tissues can be also obtained by simply using Tau and choosing the tissue with the highest expression.

Z-score and PEM score are the only methods to detect underexpression. But z-score is the most sensitive to the number of tissues used for analysis, and generally performs poorly on most tests. The PEM score performs relatively well, though it is skewed to 0, i.e. to calling genes as ubiquitous (Figure 1 and Supplementary Figure S2).

In general, almost twice as many genes can be called expressed in at least one tissue with RNA-seq than with microarray (see 'Materials and Methods' and Supplementary Figure S58). It has been reported that the detection of lowly expressed genes is better with RNA-seq than with microarrays [57–59]. Because the most tissue-specific genes are often lowly expressed [1, 4, 20], RNA-seq can detect specific genes that were not detected using microarrays (Supplementary Figure S58). We observe that the correlation between RNA-seq and microarray data set is of the same scale as the correlation between two RNA-seq data sets (Figure 6 and Supplementary Figures S55–S57). It should be noted that the correlation between microarray and RNA-seq is calculated only on half of the genes, mostly excluding specific ones, and that the second RNA-seq data set has only six tissues, which could make the correlation between RNA-seq data sets weaker.

Generally, the tissue specificity estimated from different data types appears to be different. This is notable relative to the number of tissues (Figure 2 compared with Supplementary Figure S48 and Supplementary Figure S7 compared with Supplementary Figure S49): tissue specificity calculated on microarray with a small number of tissues is poorly correlated to that with a larger number in human data, but the opposite is seen for mouse data. The correlation between species is higher for RNA-seq than for microarray (Figure 3 and Supplementary Figure S54). Our observations imply that past results, which

relied on microarray data for the evolutionary interpretation of tissue specificity, should be treated with great caution.

With any method of calculating tissue specificity, it should be noted that if the proportion of closely related tissues (e.g. different parts of the brain) in the set of tissues is high, the tissue specificity will be biased. Moreover, usually a large proportion of tissue-specific genes are testis specific, so special care should be taken in comparing data sets with and without testis. Thus, in general, during the analysis of tissue specificity, care should be taken in sampling the tissues used.

For studying the evolution of gene expression, we show here that tissue specificity is a biologically relevant parameter that has strong conservation between relatively closely related species such as human and mouse. Our results show that using a robust method such as Tau allows evolutionary comparisons even when tissue sampling somewhat differs (e.g. correlation with 27 versus 16 tissues). In light of the difficulties of comparing expression levels between species [16, 60, 61], tissue specificity holds promise not only as a confounding factor to take into account in molecular evolution [20], but also as a measure of biological function that can be compared between genes and between species.

Tissue specificity is also important for biomedical applications, as, for example, cancer malignancies can be very tissue specific [62]. More broadly, causative eQTLs identified by genome-wide association study can affect tissue-specific regulation of genes, which is linked to a weak enrichment in disease association of single nucleotide polymorphisms [63].

Conclusion

The best overall method to measure expression specificity appears to be Tau, which is reassuring, considering the number of studies in which it has been used. Counts is the simplest method, and if the threshold is chosen properly, it shows good results, although with a tendency to under-call tissue-specific genes. Gini is similar to Tau in its performance. These methods allow to capture a signal that has both functional and evolutionary significance to the genes that are studied.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/

Key Points

- Tissue specificity can be measured reliably, and carries relevant biological information.
- Tissue specificity is largely conserved between human–mouse orthologues.
- Tau is the best metric to measure tissue specificity, while Gini and simple Counts also work well.
- RNA-seq is more powerful than microarrays to detect tissue-specific genes.

Acknowledgements

We thank Marta Rosikiewicz, Iakov Davydov and Andrea Komljenovic for their helpful comments and suggestions. We thank anonymous reviewers for their constructive comments on an earlier version of this manuscript.

Funding

This work was supported by the Swiss National Science Foundation (grants number 31003A 133011/1 and 31003A_153341/1) and Etat de Vaud.

References

1. Subramanian S, Kumar S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 2004;**168**:373–81.
2. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 2000;**17**:68–74.
3. Park SG, Choi SS. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol Biol* 2010;**10**:241.
4. Lercher MJ, Urrutia AO, Hurst LD. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* 2002;**31**:180–3.
5. Vinogradov AE. Isochores and tissue-specificity. *Nucleic Acids Res* 2003;**31**:5212–20.
6. Ponger L, Duret L, Mouchiroud D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res* 2001;**11**:1854–60.
7. Liu W, Mei R, Ryder TB, *et al*. Analysis of high density expressino microarrays with signed-rank call algorithms. *Bioinformatics* 2002;**18**:1593–9.
8. Wagner GP, Kin K, Lynch VJ. A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci* 2013;**132**:159–64.
9. Hebenstreit D, Fang M, Gu M, *et al*. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* 2011;**7**:497.
10. Dezso Z, Nikolsky Y, Sviridov E, *et al*. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol* 2008;**6**:49.
11. Ramsköld D, Wang ET, Burge CB, *et al*. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 2009;**5**:e1000598.
12. Ma L, Cui P, Zhu J, *et al*. Translational selection in human: more pronounced in housekeeping genes. *Biol Direct* 2014;**9**:17.
13. Cui P, Lin Q, Ding F, *et al*. The transcript-centric mutations in human genomes. *Genomics Proteomics Bioinformatics* 2012;**10**:11–22.
14. Yanai I, Benjamin H, Shmoish M, *et al*. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 2005;**21**:650–9.
15. Smeds L, Warmuth V, Bolivar P, *et al*. Evolutionary analysis of the female-specific avian W chromosome. *Nat Commun* 2015;**6**:7330.
16. Piasecka B, Robinson-Rechavi M, Bergmann S. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics* 2012;**28**:1865–72.
17. Assis R, Bachtrog D. Neofunctionalization of young duplicate genes in Drosophila. *PNAS* 2013;**110**:17409–14.
18. Assis R, Kondrashov AS. Conserved proteins are fragile. *Mol Biol Evol* 2014;**31**:419–24.
19. Bush SJ, Kover PX, Urrutia AO. Lineage-specific sequence evolution and exon edge conservation partially explain the relationship between evolutionary rate and expression level in A. thaliana. *Mol Ecol* 2015;**24**:3093–106.

20. Kryuchkova-Mostacci N, Robinson-Rechavi M. Tissue-specific evolution of protein coding genes in human and mouse. *PLoS One* 2015;**10**:e0131673.
21. Liao B-Y, Zhang J. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol* 2006;**23**:1119–28.
22. Liao B-Y, Scott NM, Zhang J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol* 2006;**23**:2072–80.
23. Weber CC, Hurst LD. Support for multiple classes of local expression cluster in Drosophila melanogaster, but no evidence for gene order conservation. *Genome Biol* 2011;**12**:R23.
24. Zhao L, Wit J, Svetec N, *et al*. Parallel gene expression differences between low and high latitude populations of Drosophila melanogaster and *D. simulans*. *PLOS Genet* 2015;**11**:e1005184.
25. Yu X, Lin J, Zack DJ, *et al*. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res* 2006;**34**:4925–36.
26. Liu X, Yu X, Zack DJ, *et al*. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics* 2008;**9**:271.
27. Julien P, Brawand D, Soumillon M, *et al*. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol* 2012;**10**:e1001328.
28. Cortez D, Marin R, Toledo-Flores D, *et al*. Origins and functional evolution of Y chromosomes across mammals. *Nature* 2014;**508**:488–93.
29. Assis R, Bachtrog D. Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evol Biol* 2015;**15**:138.
30. Winter EE, Goodstadt L, Ponting CP. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res* 2004;**14**:54–61.
31. Schug J, Schuller W-P, Kappen C, *et al*. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* 2005;**6**:R33.
32. Vandenbon A, Nakai K. Modeling tissue-specific structural patterns in human and mouse promoters. *Nucleic Acids Res* 2010;**38**:17–25.
33. Xiao S-J, Zhang C, Zou Q, *et al*. TiSGeD: a database for tissue-specific genes. *Bioinformatics* 2010;**26**:1273–5.
34. Huminiecki L, Lloyd AT, Wolfe KH. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* 2003;**4**:31.
35. Milnthorpe AT, Soloviev M. The use of EST expression matrixes for the quality control of gene expression data. *PLoS One* 2012;**7**:e32966.
36. Russ J, Futschik ME. Comparison and consolidation of microarray data sets of human tissue expression. *BMC Genomics* 2010;**11**:305.
37. Lin H, Ouyang S, Egan A, *et al*. Characterization of paralogous protein families in rice. *BMC Plant Biol* 2008;**8**:18.
38. Divina P, Vlcek C, Strnad P, *et al*. Global transcriptome analysis of the C57BL/6J mouse testis by SAGE: evidence for non-random gene order. *BMC Genomics* 2005;**6**:29.
39. Ceriani L, Verme P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J Econ Inequal* 2012;**10**:421–43.
40. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2015.
41. Warnes G, Bolker B, Bonebakker L, *et al*. *gplots: Various R Programming Tools for Plotting Data*. R package version 2.17.0. 2015.
42. Handcock MS, Morris M. *Relative Distribution Methods in the Social Sciences*. New York, NY: Springer, 1999.
43. Handcock MS. *Relative Distribution Methods*. R package version 1.6–4. 2015.
44. Chen H. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. R package version 1.6.16. 2015.
45. Bolstad BM. *preprocessCore: A Collection of Pre-processing Functions*. R package version 1.32.0. 2015.
46. Fagerberg L, Hallstrom BM, Oksvold P, *et al*. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 2014;**13**:397–406.
47. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011;**9**:e1001046.
48. Lin S, Lin Y, Nery JR, *et al*. Comparison of the transcriptional landscapes between human and mouse tissues. *PNAS* 2014;**111**:17224–9.
49. Brawand D, Soumillon M, Necsulea A, *et al*. The evolution of gene expression levels in mammalian organs. *Nature* 2011;**478**:343–8.
50. Bastian F, Parmentier G, Roux J, *et al*. Bgee: integrating and comparing heterogeneous transcriptome data among species. *Data Integr Life Sci* 2008;**5109**:124–31.
51. Ge X, Yamamoto S, Tsutsumi S, *et al*. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* 2005;**86**:127–41.
52. Thorrez L, Van Deun K, Tranchevent LC, *et al*. Using ribosomal protein genes as reference: a tale of caution. *PLoS One* 2008;**3**:e1854.
53. Schuster EF, Blanc E, Partridge L, *et al*. Correcting for sequence biases in present/absent calls. *Genome Biol* 2007;**8**:R125.
54. Eden E, Navon R, Steinfeld I, *et al*. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 2009;**10**:48.
55. Supek F, Bošnjak M, Škunca N, *et al*. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS One* 2011;**6**:e21800.
56. Rosikiewicz M, Robinson-Rechavi M. IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinformatics* 2014;**30**:1392–9.
57. Zhao S, Fung-Leung W-P, Bittner A, *et al*. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 2014;**9**:e78644.
58. Wang C, Gong B, Bushel PR, *et al*. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol* 2014;**32**:926–32.
59. Emig D, Kacprowski T, Albrecht M. Measuring and analyzing tissue specificity of human genes and protein complexes. *EURASIP J Bioinforma Syst Biol* 2011;**2011**:5.
60. Pereira V, Waxman D, Eyre-Walker A. A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics* 2009;**183**:1597–600.
61. Gilad Y, Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research* 2015;**4**:121.
62. Maris JM, Knudson AG. Revisiting tissue specificity of germline cancer predisposing mutations. *Nat Rev Cancer* 2015;**15**:65–6.
63. Göring HHH. Tissue specificity of genetic regulation of gene expression. *Nat Genet* 2012;**44**:1077–8.