Evaluation of diversity, specialization, and gene specificity in transcriptomes

Octavio Martínez* and Humberto Reyes-Valdés**

*Laboratorio Nacional de Genómica para la Biodiversidad (Langebio) Cinvestav Irapuato, México. ** Department of Plant Breeding, Universidad Autónoma Agraria Antonio Narro, Saltillo, Coahuila, México

The transcriptome

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.

Transcriptome Diversity

Consider the division of an organism in tissues, then the transcriptomes of each tissue can be simply described as the set of relative frequencies, p_{ij}, for the *i-th* gene (*i = 1, 2, …, g*) in the *j-th* tissue (*j = 1, 2, …, t*). Then the diversity of the transcriptome of each tissue can be quantified by an adaptation of Shannon´s entropy formula (1),

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

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: $1/g$.

Gene specificity

Consider the average frequency of the *i-th* gene among tissues, p_i, then gene specificity is given by

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

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.

Relative Transcriptome Specialization

Tissue specialization can be measured by the average of the gene specificities, say,

$$\delta_j = \sum_{i=1}^{g} p_{ij} S_i$$

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

Fig. 1 presents an example of the point estimation of transcriptome diversity and specialization in 36 human tissues, derived from microarray data.

Statistical properties of the estimators

Assuming a multinomial distribution for the number of gene tags in the transcriptome it is possible to give approximate values for the expectation and variance of the estimator of transcriptome diversity (H_j). These expressions, found by Basharin (2), are given by

$$E[\hat{H}_j] = H_j - \frac{g-1}{2n_j}\log_2(e) + O(\frac{1}{n_j^2})$$

$$V[\hat{H}_j] = \frac{1}{n_j}[\sum_{i=1}^{g} p_{ij}(\log_2(p_{ij}))^2 - H_j^2] + O(\frac{1}{n_j^2})$$

where E[•] and V[•] represent the expectation and variance operators respectively, O(1/n^2) designates a quantity of order n_j^{-2} and the constant $log_2(e)$ has a value of 1.442695 and results from the use of logarithms base 2 instead of natural logarithms. From the first equation it is apparent that the estimator is biased, and for values of g larger than n_j the estimator will seriously underestimate the transcriptome diversity.

The expectation and variances of the gene specificities, *Si* as well as transcriptome specialization, δj, are too complex to be obtained analytically. However they can be assessed by re-sampling the original distribution to obtain approximate but robust inferences.

Approximate confidence intervals for the information parameters

It is possible to obtain confidence intervals for the transcriptome diversity (H_j) by using the asymptotic normal distribution of the estimator (2), and the same can be achieved for this parameter as well as for the transcriptome specialization (δ_j) and gene specificities (S_i) by re-sampling the gene tag datasets assuming the multinomial distribution (3).

The confidence interval methods tested were: 1) Asymptotic Interval or AI, assuming the normal distribution and using the approximation for the variance of H_j; 2) using the mean and variances obtained from the bootstrap procedure and assuming normality, that we name here Bootstrap Asymptotic Interval or BAI; 3) using the Bootstrap Percentiles Interval or BPI and finally 4) The method proposed by Hall (4) based in re-sampling the error of the estimation and named here as Hall Bootstrap Percentiles or HBP.

To test the methods for the estimation of confidence intervals we simulated three datasets of sample sizes n_j = 100, 1000 and 10000, each one constituted by four transcriptomes (t=4) and 16 genes (g=16). The values of p_{ij} were selected to have values of H_j of approximately equal to 1, 2, 3 and 4 (the maximum) and also maximum diversity in specialization among the transcriptomes. A total of B=10,000 bootstrap replicates were obtained from each dataset.

From Fig 2 we can observe that for the maximum value of H_j = 4 both, the BAI and BPI intervals fail to include the true value of the parameter, and the HBP fall to the right of the parametric value overestimating it out of its possible rank (obvious in sample size 100). In all the other cases, except for the value of H_j = 1 with n_j = 100, the true value of H_j is included within the intervals. This means that when the sample size is small all methods could fail to include the true value of H_j. The differences between the lengths of the intervals obtained by the three methods (AI, BAI and BPI) for H_j for the same sample size are not large. However, given the lack of normality that could be present when the sample size is relatively small, it appears safer to use the bootstrap percentile (BPI or HBP) to obtain confidence intervals for this function.

From Fig. 3 we can notice that the majority of the intervals, for all sample sizes, include the true value of the parameter, except for the case of the gene 4 for which all intervals for the sample size n_j=1000 fails to include the true value. For all cases studied the specificity estimator is positively biased, i.e., it tends to overestimate the specificity of the gene. The Shapiro-Wilks test for normality of the bootstrap replicates of S_i demonstrated that in the majority of the cases its distribution is not normal, and thus for this case the BPI must be preferred over the other methods to obtain confidence intervals.

From Fig. 4 we can see that the true value of δ_i is included in all the HBP intervals (limits shown as diamonds in the graph), however, for some cases the BAI and BPI intervals fail to include the true value of δ_1. In the case of δ_i the HBP intervals are much better centered on δ_i than the BAI and BPI intervals. On these grounds the HBP method can be preferred for this function. We can also notice that the estimator of δ_i is consistently biased to the right of the true parameter. The Shapiro-Wilk test applied to the bootstrap replicates of showed that normality can be assumed for the estimator of δ_i only for large sample sizes.

Example of estimation of transcriptome properties of human organs

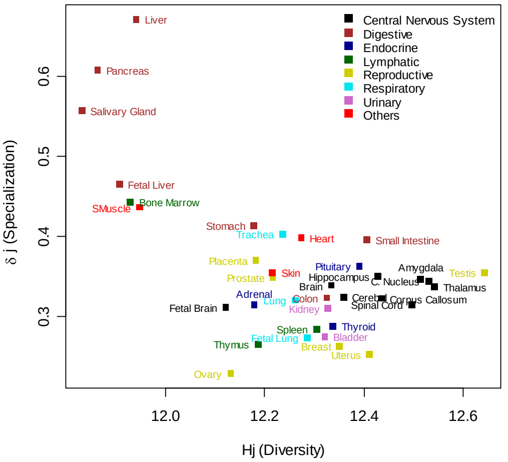

Fig. 1 Scatter plot of estimated values of *Hj* (diversity) *versus δj* (specialization, given by the average gene specificity) for 36 tissues of the human systems from microarray data, GEO accession GDS1096. Tissues are coloured by system of origin.

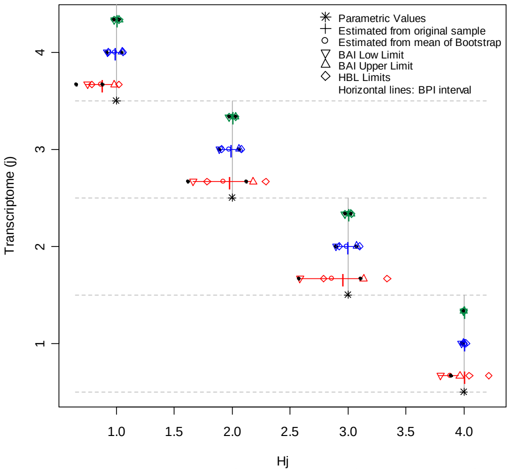

Fig. 2 Approximated 95% confidence intervals for the values of H_j. Transcriptomes (*j*), are presented in the Y axis separated by gray discontinuous lines. The scale in the X axis is for the values of H_j. Intervals are colored by sample size; red for n_j=100, blue for n_j=1000 and green for n_j=10,000. Parametric values of H_j for each transcriptome are represented by black asterisks over the grey discontinuous lines. Values of the estimated H_j are represented by colored vertical lines, while the values of the means of the B=10,000 bootstrap replicates are shown as open circles. The limits for the BAI intervals are denoted by colored triangles pointing below (lower limit) or above (upper limits), AI limits are shown as black points while the BPI intervals are shown as continuous color lines and HBP limits are shown as red diamonds.

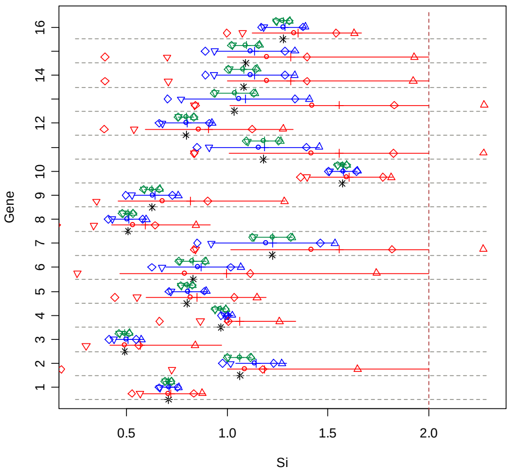

Fig. 3 Approximated 95% confidence intervals for the values of gene specificities, S_i. Genes in Y axis. Colors and annotations as in Fig 2.

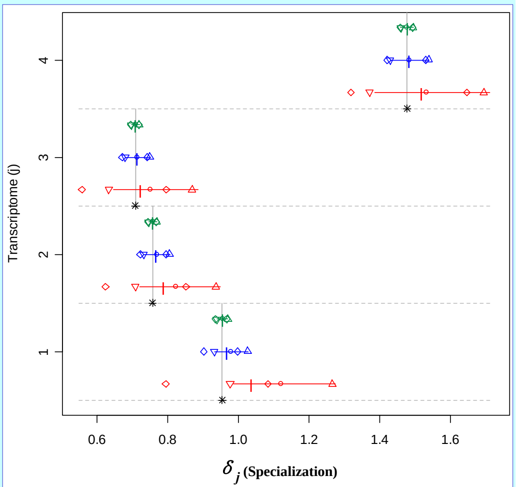

Fig. 4 Approximated 95% confidence intervals for the values of δ_j. Colors and annotations as in figures 2 and 3.

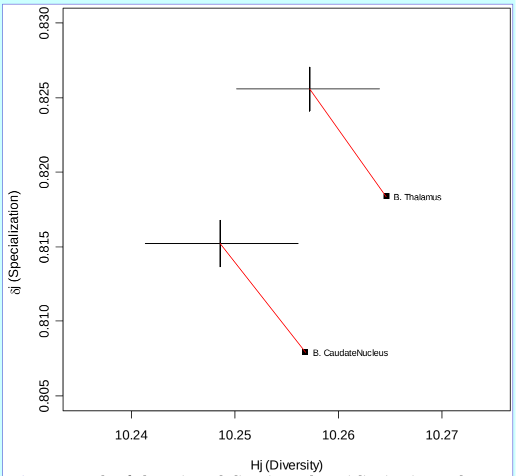

Fig. 5 Example of the estimated diversity and specialization in two human tissues from an analysis of MPSS data (1). Squares are the values originally estimated in the sample. Black lines are 95% approximate confidence intervals for the parameters using the BPI method. The intervals are positioned into the mean of the bootstrap replicates to show the bias in both functions (Hj and δ j). A red line joins the estimates in the original sample and the point where the means of the bootstrap replicates are.

Fig. 5 shows, with real data, how the correction for bias and the estimation of confidence intervals are fundamental for the correct inference of the properties of the transcriptomes.

Conclusions and Perspectives

The methods presented here allow performing correct inferences about the diversity and specialization of the transcriptomes as well as about the specificities of the genes that are expressed in them. For transcriptome diversities (H_j) and gene specificities (S_i), the bootstrap percentile method (BPI) appears to be a reasonable solution for the confidence interval estimation, while for the transcriptome specialization (δ_j) the Hall bootstrap percentile (HBP) appears to be a better choice. The methods shown are easily extended to perform test of hypothesis for the parameters of interest; this can be done by obtaining bootstrap estimates of the relevant difference and calculating the BPI. If the BPI contains zero (at a given confidence level) then the null hypothesis is not rejected. This permits to decide, for example, if two genes are equally specific for a set of transcriptomes or if two transcriptomes differ significantly in diversity or specialization. A lot of biologically relevant questions can be answered in this way.

The methods presented are only applicable to the counting of gene tags, as in SAGE, MPSS, large collections of ESTs, pirosequencing (454) or any other high throughput sequencing strategy. However, these methods are not directly applicable to continuous measurement of gene expression, as the ones performed in microarrays.

References

1. Martínez, O. & Reyes-Valdés, H. (2008) *Defining diversity, specialization, and gene specificity in transcriptomes through information theory*. PNAS 105, 9709–9714.

2. Basharin, G. P. (1959) On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. Theory of Probability and its Applications 4, 333-336.

3. Efron, B. & Tibshirani, R. J. (1993) An introduction to the bootstrap (Chapman & Hall, New York - London).

4. Hall, P. (1992) The Bootstrap and Edgeworth Expansion (Springer, New York).

Acknowledgments

We are thankful to Conacyt, Concyteg, Cinvestav and UAAAN for the support given to this project.

4th. EMBO Conference *From Functional Genomics to Systems Biology*, EMBL Heidelberg, 15 - 18 November, 2008.