

***In Silico* Identification and Analysis of Tissue-Specific Genes using the Database of Human Expressed Sequence Tags**

Sheng-Ying Pao^{1, 2}, Win-Li Lin¹, Ming-Jing Hwang²

Keywords: tissue-specific genes, genome wide expression profile, EST

1 Introduction.

Genome wide transcriptome analysis with histological information can provide insights to identify candidate genes that are differentially expressed in certain tissues. At the transcriptome level, differential expression of genes plays key roles in maintaining and regulating cellular functions. Genes preferentially expressed can be characterized by their significantly different expression levels of transcripts in various tissues. In this study, tissue-specific genes were identified using human expressed sequence tags (EST).

2 Methods.

EST data, GenBank reports from dbEST, and UniGene build #161 were downloaded from NCBI (National Center for Biotechnology information). We extracted a description triplet for each EST library under the field “Title” (Lib. Name), “Tissue” and “Organ”. According to the triplet, each library was classified into a corresponding category of tissue. The classification process is illustrated in fig.1 wherein libraries are automatically classified by tissue and organ unless they are different, both null, or inconsistent with the title; in these cases, manual classifying and checking with title was carried out. To mitigate variation due to unspecified tissue and artificially modified expression, libraries described as subtracted, differentially displayed, normalized, or coming from multiple tissues were excluded. Libraries without a clear description in the triplet were also discarded. By subjectively selecting category names of classified libraries, a “target_tissue list” consisted of interested tissue names was then generated.

For each category in the “target_tissue list”, the corresponding EST gi numbers and the UniGene clusters to which they belong were retrieved automatically. For each target tissue in the “target_tissue list”, we performed the differential expression evaluation according to [1]:

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!y! \left(1 + \frac{N_2}{N_1}\right)^{(x+y+1)}}$$

where x and y are the number of ESTs matching the UniGene cluster from the target tissue and from all the other tissues, respectively. Likewise, N₁ and N₂ are the total number of ESTs from the target tissue and from all the other tissues, respectively. Clusters with N₁<1000, N₂<1000, x 0.05*N₁, or y 0.05*N₂ were excluded from the test for insufficient sample size.

¹Institute of Biomedical Engineering, National Taiwan University College of Medicine, No.1 Jen-Ai Road Section 1, Taipei, Taiwan. E-mail: r91548018@ntu.edu.tw

² Institute of Biomedical Sciences, Academia Sinica, No.128 Yen-Chiu-Yuan Road Section 2, Taipei, Taiwan. E-mail: mjhwang@ibms.sinica.edu.tw

3 Results and discussion.

We discarded 1898 EST libraries as described in methods, leaving 6,247 libraries with 3,352,546 ESTs for analysis. Along with UniGene build #161, they provide sufficient data to identify tissue specific genes by *in silico* data mining. For these libraries we constructed a tissue hierarchy and classified the libraries into 343 categories of tissues, of which we selected 113 categories to form a target_tissue list to detect differentially expressed genes. This “target_tissue list” provides a flexible framework for detecting differential expression in tissues of interest for various research purposes aiming to, for example, analyze the expression divergence of genes in one special cell type from multiple tissues, or to compare genes specific to diseased and normal states.

To evaluate our approach, we compared the placenta-specific genes identified in our results with those of Ref [2], in which 90 preferentially expressed genes were reported, of which 19 have subsequently been removed from UniGene and have become nonexistent in UniGene build #161. Our analysis yielded 508 genes preferentially expressed in placental libraries with $p < 10^{-6}$, which included all of the 71 genes reported in [2] that remained existent in UniGene build #161.

In summary, we have constructed a tissue hierarchy for EST libraries and identified differential expressed genes in 113 tissues. The tissue-specific gene information derived from this study will be useful in functional genomics research.

4 Figures.

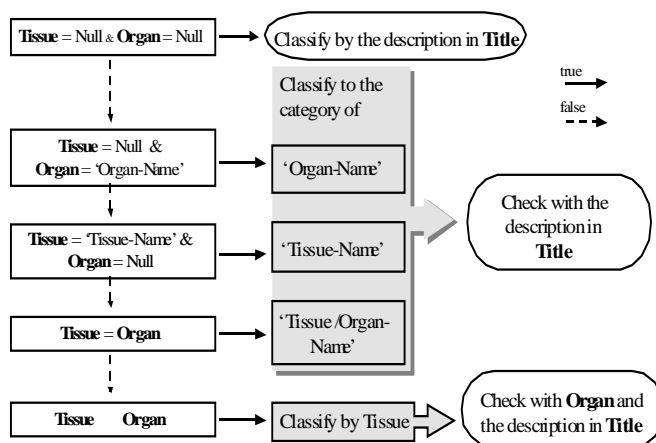


Figure 1. Procedure for EST library classification.

References

- [1] Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Research.*, 7: 986-995.
- [2] Miner, D. and Rajkovic, A. 2003. Identification of expressed sequence tags preferentially expressed in human placentas by in silico subtraction. *Prenatal Diagnosis.*, 23(5):410-419.