Global coexpression analysis of human protein-coding genes

Katia de Paiva Lopes[1,2], Francisco José Campos-Laborie[2], Ricardo Assunção Vialle[1], José Miguel Ortega[1], Javier De Las Rivas[2]

[1]*Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais (UFMG), Brasil.*

[2]*Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Cientificas (CSIC), Salamanca, Spain*

Advances in high throughput sequencing technologies have introduced a new alternative to transcriptome analysis, namely RNA-Seq. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptome. However, until recently, little has been reported about the determinants of human cell identity, particularly from the joint perspective of gene evolution and expression. In view of these, our work presents a combined analysis of human transcriptome data: 1) An evolutionary analysis using a RNA-Seq dataset of 116 samples from 32 tissues (E-MTAB-2836) using a database of orthologous proteins (OMA); and 2) a relational context of the human protein-coding genes based on a robust coexpression network analysis. Therefore, we present a complex network –like a galaxy– that includes 1,691 protein nodes related with 19,615 interactions. This network corresponds to a subset of the coexpression network, which includes 2,298 proteins and 20,005 interactions. The coexpression dataset was built calculating the pair-wise Spearman correlation coefficient (r) of all the genes along the 116 samples and only selecting, as positive gene-pairs, the ones that had a correlation coefficient ≥ 0.85. A cross-validation of these correlation values was also applied by a random selection of two sample replicates from each tissue (i.e. a total of 32 x 2 = 64 samples) and recalculating again the Spearman correlation for these random subsets of the data. This sampling was done 100 times, annotating for each gene-pair the number of times that its r coefficient was ≥ 0.85. Only the gene-pairs validated 100 time in this sampling were selected. The analysis of the network done with MCODE revealed the existence of 11 major subnetworks –considered as major constellations in the galaxy of nodes– that had a clear enrichment in certain groups or modules of highly coexpressed proteins showing a tendency to include proteins of the same evolutionary age. Finally, the study of the pair-wise correlation of the gene expression profiles along tissues allowed building human gene coexpression networks and find modules with functional and biological meaning where we did map the age of the genes and demonstrate the existence of tighter links between age-related proteins.

Supported by: Capes, FAPEMIG, CNPq.