Comparative analysis of transcriptomes reveals the existence of genes with distinct profiles: overactive genes and gaussian genes

Lissur Azevedo Orsine[1], Glaura da Conceição Franco[2], José Miguel Ortega[3],

1 UFMG
2 Departamento de Estatística, Universidade Federal de Minas Gerais
3 Universidade Federal de Minas Gerais. Laboratório de Biodados

Abstract

Presently several experiments of characterization of transcription in different tissues have been conducted. Here we analyzed the expression profiles of five experiments: ENCODE, FANTOM5, GTEx, Illumina and Uhlen, comprising respectively, 13, 56, 53, 16 and 32 tissues. They express over 1e-6 TPM (transcripts per million), respectively, 42, 21, 57, 47 and 44 thousands of transcripts. We have noticed that some genes vary the expression level around a mean value. Applying a statistic test, we determined that 13%, 10%, 3%, 18% and 12% of genes pass a test for normal distribution, thus we refer to these as primarily Gaussian (pG) genes. We also observed that often a gene has a high expression in a couple of tissues, but the remaining ones show a normal distribution. Removing the outliers (obtained from the boxplot procedure), we depicted cases where in some tissues the gene is overactive (in relation to the term "overexpressed") and in the complementary tissues the gene behave as Gaussian. Therefore, in some tissues the gene was considered complementary Gaussian (cG) while in other tissues the gene was overactive. The percentage of total Gaussian genes, adding up pG and cG increased to 68%, 21%, 11%, 60% and 61%. Thus, a regulatory mechanism that produces an average expression is very common. Moreover, overactive genes may play an important role in the tissues where they are overactive. A third category of genes was depicted by analyzing those that did not surpass a threshold of expression in the third quartile, a metrics derived from the boxplot procedure. These genes were named Tissue Specific (TS), because the expression is low in at least 3/4 of the set and where they are expressed, they present sharp peaks. Respectively, the percentage was 35%, 22%, 48%, 27% and 36%. We noticed that after removing outliers, actually the complementary tissues in some cases may present a very low but Gaussian expression. The percentage of genes that are TS in some tissues and the complementary set of tissues show Gaussian expression, represent, respectively, 35%, 0%, 1%, 27% and 36% of the total genes. The broad variation of these proportions is probably due to the accuracy of determining expression near the background by the distinct methodologies. However, the expression around an average is a remarkably frequent feature rather than an exception. Further analysis of the composition of the reported categories will be presented.

Funding: CAPES, CNPq and FAPEMIG