Systematic Normalization with Multiple Housekeeping Genes for the Discovery of Genetic Dependencies in Cancer

Oliver Bonham-Carter and Yee Mon Thu
Depts of Computer Science and Biology, Allegheny College, Meadville, PA

https://www.cs.allegheny.edu
obonhamcarter@allegheny.edu
ythu@allegheny.edu

PROJECT OBJECTIVES

We analyze gene expression data to discover pairs of genes whose expressions demonstrate patterns of correlation using a computational approach. This project presents:

► A focus on genes suppressing genome instability (GIS genes) since function or expression may often be altered in cancer.
► A computational method to determine normalizing factors that make it possible to discover pairs of GIS genes which show consistent correlation.
► Normalizing factors, created by a selection of cancer-specific housekeeping genes providing ability to compare gene expressions data and to treat these values by linear regression.

GENE CORRELATION

Genetic interactions between genes are often responsible for DNA repair or genome stability.

► FPKM datasets are used to determine the existence of a positive or negative correlation between the expression of two given GIS (suppress genome instability) genes in cancer, Shown in Figure 2.
► Correlations can reveal if two GIS genes coordinate or if an alteration in the expression of one GIS gene increases dependency of cancer cells on another GIS gene.

METHOD: PROOF OF CONCEPT

Random selection of ten data sets of breast cancer gene expression.

► R² values in heatmaps from linear regression models, all-against-all regressions of GIS genes.
► Left to right, R² values, derived from single housekeeping genes to create normalizing factors; HKG_Tubb, HKG_Tuba1a, HKG_Tubb, HKG_AvgTubbTuba1aGapdh, see Figure 5,(A-C), resp.
► Ten housekeeping genes: HKG_AvgG1, HKG_AvgG2, HKG_AvgG3, See Figure 5,(D-F), resp.

RESULTS

► HKG_Tuba1a (B): high diversity of R² values indicated poor normalizing, biologically improbable.
► According to regression model results, normalizing factors that were created from larger groups of housekeeping gene produced correlations implying more biological relevance.

CONCLUSIONS

► Our results (examples shown in heatmaps of Figure 5) indicate normalized data and contain biologically probable findings.
► Single Expression Normalization: Normalizing factors derived from single housekeeping genes did not provide generally consistent correlations that were biologically probable.
► Multiple Expression Normalization: Using the averaged expression values of multiple housekeeping genes was an effective approach to finding biologically relevant consistency across our data sets.
► Our method enabled us to identify the co-expression of gene pairs in breast cancer tissues and compared diverse normalization factors.
► Our study also allows reproducibility across data sets and allows for scalability in gene correlation research.