

The dynamic range of gene expressions depend on their ontology

Yizhou Xie

*Max McGee National Research Center for
Juvenile Diabetes,
Medical College of Wisconsin,
Milwaukee, WI 53226, USA
yxie@mail.brc.mcw.edu*

Soumitra Ghosh

*Max McGee National Research Center for
Juvenile Diabetes, and Human and
Molecular Genetics Center,
Medical College of Wisconsin,
Milwaukee, WI 53226, USA
sghosh@mcw.edu*

Parthav Jailwalia,

*Max McGee National Research Center for
Juvenile Diabetes,
Medical College of Wisconsin,
Milwaukee, WI 53226, USA
pjailwal@mail.mcw.edu*

Xujing Wang

*Max McGee National Research Center for
Juvenile Diabetes, and Human and
Molecular Genetics Center,
Medical College of Wisconsin,
Milwaukee, WI 53226, USA
xujing@mcw.edu*

Abstract

In this study we investigate the correlation between the normal variation of gene expressions and their ontologies. Three sets of time-series microarray data are utilized that represent three different but typical physiological/developmental processes. We analyze the dynamic range and the variation of expression levels during these processes for genes in each of the top-level ontologies of the biological process, molecular function and cellular component, as defined by GO. Significant and consistent differences are observed. For example, transcription factors were found to have consistent narrower dynamic range of expression in all three processes. The findings will shed light on study design, statistical evaluation, and data mining of genetic data including microarray data.

1. Data preparation

Comprehensive profiling of gene expressions during physiological and/or pathological processes is becoming a standard research approach; yet quantitative interpretation of data is still difficult. This problem is mainly due to the current poor understanding of the intrinsic variation in the

expression level of individual genes. As a result, we do not have a priori knowledge of the expression distribution for each gene, and we have to use the average behavior of all genes on a microarray slide to derive the statistics for each individual gene. This can be problematic, as it is known that naturally some transcripts are present at relatively constant levels, while others are expressed at highly variable levels. However, a comprehensive characterization of the gene-dependent expression variability is still missing.

We approach this issue by the examination of the dynamic range of expressions of different gene ontology (GO) categories. To capture the natural fluctuation in transcript levels we decided to profile the basic dynamic process for each cell population: proliferation, programmed cell death, and cell differentiation. We have identified three datasets that profiled expression changes during these processes: (1) the yeast cell cycle data from Stanford University (<http://genome-www.stanford.edu/cellcycle/>) [1]. Gene expressions were profiled throughout cell cycle, after synchronized by three independent methods: α factor arrest, elutriation, and arrest of a *cdc15* temperature-sensitive mutant; (2) the pancreas perinatal developmental data from the EpconDB (<http://www.cbil.upenn.edu/EPConDB/>), where, pancreas gene expression changes were profiled during its developments, at embryonic days 14.5, 16.5 18.5,

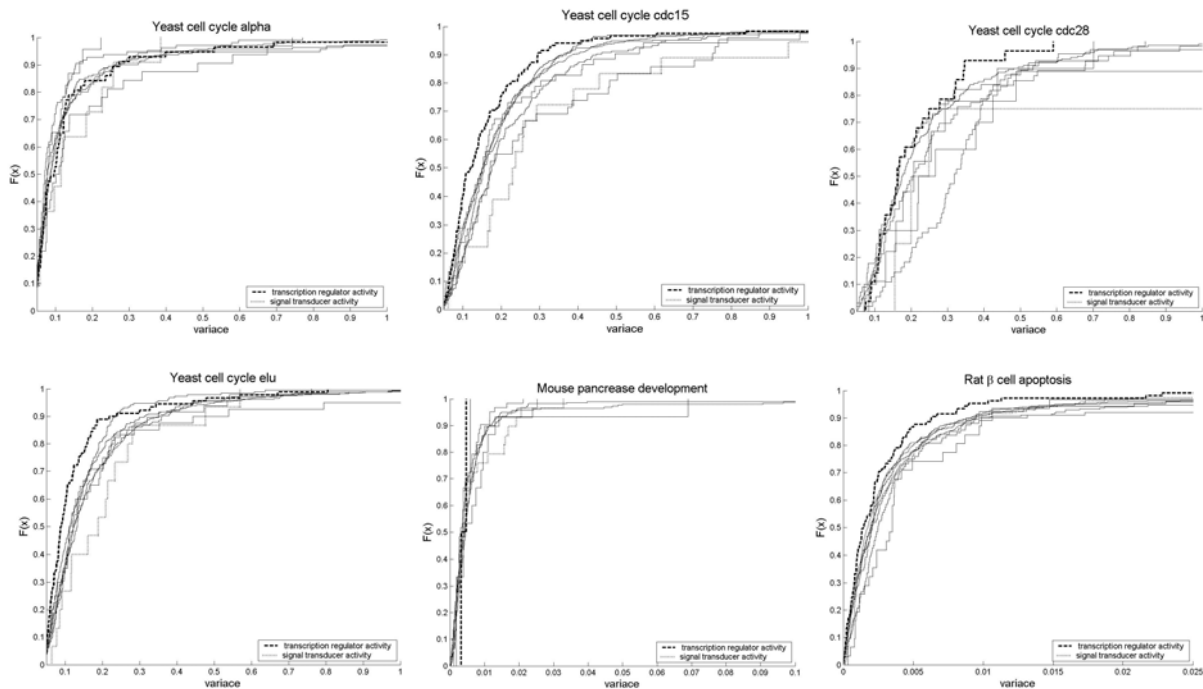


Figure 1. Cumulative fraction plots for variance distribution.

birth, 7 day, and adulthood; and (3) the apoptosis progression of pancreatic islet β -cells data from our own lab. Islet RIN-5mF cells were treated with staurosporine at $1\mu\text{M}$ for 0, 2, 4, and 6 hours, and gene expression changes were profiled.

For each dataset, we bring in top-level GO annotations under biological process, cellular component, and molecular function. For example, under molecular function, the categories exist for all three datasets include: binding, structural molecule activity, catalytic activity, transcription regulator activity, chaperone activity, transporter activity, enzyme regulator activity, signal transducer activity. We then calculate the variance of gene expression across the time course for genes in each category.

2. Results

The variance shows a clear dependence on their ontology. Figure 1 gives the cumulative fraction plot of the variance distribution for the molecular function categories from three experiments. We have specifically labeled the transcription regulator activity, and the signal transducer activity. Evidently they show a consistent difference from the rest categories. Distribution for genes involved in transcription regulator activity tend to be skewed toward the smaller variation, while distribution for signal transducer

activity usually skewed toward the high variance. These differences should be incorporated in the statistical evaluation of microarray data. For example, a same fold of change in transcription factors can be more significant than in signal transducer activity genes.

On top of the general agreements, difference exists among the datasets. Some can be explained by the specific characteristics underlying each study design. For example, the mouse pancreas dataset consist of significantly fewer genes than the others (3,840 clones, versus 6,178 for the yeast data, and 18,432 for the rat data); the apoptosis is drug-induced, which may not represent well the in vivo programmed cell death process under normal physiological conditions. Nevertheless, some difference can be intrinsic to the process under investigation, and may be utilized to further characterize the process.

3. References

- [1] Spellman, P.T., et al., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*, 1998. 9(12): p. 3273-97.