

Extra Views

# Gene Expression Phenotypes of Oncogenic Signaling Pathways

Erich S. Huang<sup>1,3</sup>

Esther P. Black<sup>1,3</sup>

Holly Dressman<sup>1,3</sup>

Mike West<sup>2,3</sup>

Joseph R. Nevins<sup>1,3,4,\*</sup>

<sup>1</sup>Department of Molecular Genetics and Microbiology; Duke University Medical Center; <sup>2</sup>Institute of Statistics and Decision Sciences; <sup>3</sup>Computational and Applied Genomics Program; Duke Institute for Genome Sciences and Policy; Duke University; Durham, North Carolina USA

<sup>4</sup>Howard Hughes Medical Institute; Durham, North Carolina USA

\*Correspondence to: Joseph R. Nevins, Ph.D.; Department of Molecular Genetics and Microbiology; Duke University Medical Center; 366 CARL Building; Box 3054 DUMC; Durham, North Carolina 27710 USA; Tel.: 919. 684.2746; Fax: 919.681.8973; Email: j.nevins@duke.edu

Received 07/22/03; Accepted 07/23/03

Previously published online as a Cell Cycle E-publication at:  
<http://www.landesbioscience.com/journals/cc/to/new25.php?volume=2&issue=5>

## KEY WORDS

microarray, metagenes, oncogenic signatures, tumor phenotypes, Myc, Ras, Rb, E2F

## INTRODUCTION

The study of oncogenic signaling pathways has progressed remarkably over the past few decades, resulting in the identification of a large collection of activated receptors, receptor-coupled activators, kinases, phosphatases, transcription factors, and various negative regulators of these activities.<sup>1-3</sup> Nevertheless, most of the details regarding the pathways regulated by these oncogenic proteins is limited to events linearly associated with the pathway—by their nature, the discoveries have been incremental, adding new information piece by piece. Now, however, with the advent of technologies that have the power to greatly expand the scope of such studies through genome-scale analyses of gene expression patterns, large-scale analysis of patterns of protein accumulation, protein modification, and protein interaction, we have the potential to rapidly expand the understanding of these pathways and, most importantly, to integrate the action of the pathways into a unified picture.

Our recent study<sup>4</sup> represents a first step towards this goal of building a broader understanding of oncogenic signaling events. This work aims to lay a foundation for using the complex, massively multivariate data of genome-scale expression analyses for identifying and quantifying fundamental physiologic states in cell proliferation as well as pathologic states in neoplasia. Although the application of DNA microarray technology for the analysis of gene expression has been applied in numerous studies, including many focused on the analysis of oncogenic pathways,<sup>5-8</sup> the focus and concepts underlying our analytic methods define an approach that differs from past studies and that offers the potential to broaden the understanding of oncogenic pathways. DNA microarrays have the capacity to score the activity of thousands of genes simultaneously—in one sense, providing a very high-throughput northern analysis. For a molecular biologist, analyses of gene expression are all about fold change—a gene induced 10-fold is often considered more important than a gene induced 2-fold. Indeed, invoking statistics in a molecular biology experiment to ascertain the significance of a 2-fold change is simply not done—the observed change is either visually clear or not and no level of statistical significance will alter this conclusion. Genomic technology in the form of DNA microarray data has changed all of this dramatically. The use of DNA microarray data is not about the simple measures of differential expression gene-by-gene. Rather, the power lies in the ability to assay many thousands of genes simultaneously and evaluate the multivariate patterns of change across subsets that characterize a physiological or clinical state.

This complexity opens the way to powerful tools of statistical analysis—not merely simple measures of reproducibility but identification of complex patterns within the data that reflect biology. In a very real sense, this approach is no different than the analysis of complex financial datasets to identify underlying social or political factors that influence currency valuation. The power of genomic technology, generating datasets of enormous complexity, heralds the transformation of biology into a quantitative science.

## APPLICATION OF GENE EXPRESSION TECHNOLOGY TO ANALYSIS OF ONCOGENIC PATHWAYS

Our goal in this work was two-fold. First, to use the power of gene expression analysis to uncover greater detail in these pathways. Second, to use the analysis to define ‘signatures’ of the activation of the oncogenic pathways that could then be applied to the study of tumor development. The process began first with using a controlled experimental system for inducing activities known to be involved in proliferation and oncogenesis in mouse embryo fibroblasts (MEFs). We subsequently collected gene expression information from these cells, identifying structure in these data. Using robust statistical procedures to correlate data structure (termed here “metagenes”) with the state of pathway deregulation within the cells, we developed testable models from these experimental data that are able to:

1. identify the presence of these activities in external and independent validation datasets;
2. quantify the level of such activities in validation experiments; and
3. robustly distinguish such activities in an *in vivo* setting.

We collected data from a series of carefully-controlled experiments involving deregulated expression of Myc, Ras, and E2Fs -1, -2, and -3. We then statistically identified groups of genes, or “metagenes” that were highly predictive of the presence of Myc, Ras, or the E2Fs, and tested these metagenes with independent validation datasets including cell cycle and tumor data. We found that metagenes elicited by these activities in an experimentally-induced setting are reflective, and in fact predictive, of these same activities in both normal cell proliferation as well as in neoplasia.

Essential to our analytic strategy was ensuring that our interpretation of experimental data was not subject to statistical over-fitting, a common issue when analyzing large and inherently “noisy” datasets. When a group of experiments encompasses more than one million data points, random fluctuations in gene expression are easily mistaken as scientifically significant. Thus, the metagene models are tested both by out-of-sample cross-validation and against external data. The process of validation is simple: test whether an analysis is genuine by assessing its ability to predict the state of a new sample. The preliminary test uses cross-validation—one sample of the group is removed and the analysis is carried out to completely regenerate the model excluding that sample. The newly generated model is then used to test that sample as a new “unknown”. Therefore, if one has twenty replicates for one experimental condition, the process is repeated twenty times, and the model is regenerated twenty times; in each case, nineteen samples are used to generate a model. This represents a truly ‘honest’ prediction in which the sample being analyzed is not used for the generation of the predictor. In the case of a Myc metagene model, it is the regenerated model that “decides” whether the held out sample is a Myc experiment (an outcome scored as 1) or a Control experiment (scored as 0). If the regenerated model is “unsure” of the status of the held out sample it will deliver an intermediate score such as 0.5. If a model appears suitably stable in cross-validation it is applied to an independent set of data.

We took the evaluation further by assessing predictions against independently generated data—prediction of activity of the pathways during a cell cycle and prediction of tumors arising from deregulation of Myc or Ras. This latter test represents a more challenging and ultimately more rigorous validation of a metagene model’s predictive capabilities since the mouse mammary tumor model comprises not only an entirely different tissue type than fibroblasts, but also exhibits greater heterogeneity contributed by vascular, inflammatory, and stromal tissue. This also represents an application that highlights the potential for such work to truly improve our understanding of oncogenesis. We applied both Myc and Ras metagene models to mammary tumors from MMTV-Myc and MMTV-Ras transgenic mouse models. The results show that Myc and Ras models “trained” by *in vitro* experiments accurately predict whether tumors were Myc- or Ras-induced *in vivo*. Further, MMTV-Myc tumors that were confirmed by sequencing to possess sporadic k-Ras mutations were also clearly predicted by the models to possess Ras activity.

Metagene models treat physiologic and pathologic states as composites of numerous co-varying, but not necessarily coregulated genes. An important component of the metagene concept is that coefficients are assigned to individual genes in a model, allowing them to be prioritized by their impact in a model and to provide the mathematical basis for quantifying the level of a complex activity, thus generating initial insights and suggestions for gene-specific follow-on studies.

## FUTURE APPLICATIONS AND CHALLENGES

We believe the general strategy that we have outlined in two recent studies<sup>4,9</sup> is applicable in a variety of settings. Most obvious is the analysis of human cancer—the ability to identify deregulation of an oncogenic signaling pathway by virtue of detecting the ‘signature’ of the pathway. In a sense, this is an extension of the previous studies of Vogelstein and colleagues that identified genetic alterations in developing colon cancer. By using gene expression signatures rather than specific gene mutations, we detect the consequence of the mutation in the form of pathway deregulation, irrespective of how the pathway might have been altered. Thus, even if the known oncogene is not mutated, but rather another component of the pathway is altered, a model based on gene expression profiles will still detect the alteration. Building a library of gene expression signatures of oncogenic pathways we believe will facilitate characterization of human tumor development with respect to deregulated pathways.

This strategy, of recognizing a gene expression signature of a biological event is equally applicable to the study of response of cells or tissues to growth regulatory ligands, hormones, or toxins and drugs. The power is the use of multiple gene expression values—so-called metagenes – rather than individual genes or proteins as biomarkers. The same concepts underlie the development of gene expression signatures in predictive models, such as our own work in breast cancer,<sup>10,11</sup> and the merging of novel metagene patterns into such clinically oriented studies is one additional important goal, and a key current objective as we refine the statistical analysis methods that isolate and define these patterns from multiple sources of data, both experimental and observational.

We believe the next critical step in the application of these methodologies to the study of cellular signaling pathways will be the ability to integrate the gene expression signatures in a way that allows an understanding of how regulatory signaling pathways operate together, in synchrony. How does the activity of one pathway influence the signature of another? Is it possible to detect the subtle but undoubtedly significant influences of pathway interaction and synergy in the ultimate determination of cellular phenotype?

Finally, we also recognize that there are major challenges in reaping the full potential of this data analysis. The most immediate is interpreting the meaning of the sets of genes that are identified in the profiling experiments. Obviously, the gene expression profiles are not just diagnostic tools—they also represent a window into the underlying biology. Nevertheless, the capacity to interpret the meaning of the genes recovered in such a profile, and then to translate this into a better understanding of the biological processes, is severely limited by our current state of biological knowledge.

### References

1. Hunter T. Oncoprotein networks. *Cell* 1997; 88:333-46.
2. Hanahan D, Weinberg RA. The Hallmarks of cancer. *Cell* 2000; 100:57-70.
3. Sherr CJ. Cancer cell cycles. *Science* 1996; 274:1672-7.
4. Huang E, Ishida S, Pittman J, Dressman H, West M, Nevins JR. Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet* 2003; 34:226-30.
5. Collier HA, Grandori C, Tamayo P, Colbert T, Lander ES, Eisenman RN, Golub TR. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc Natl Acad Sci USA* 2000; 97:3260-5.
6. Watson JD, Oster SK, Shago M, Khosravi F, Penn LZ. Identifying genes regulated in a Myc-dependent manner. *J Biol Chem* 2002; 277:36921-30.
7. Guo QM, Malek RL, Kim S, Chiao C, He M, Ruffly M, Sanka K, Lee NH, Dang CV, Liu ET. Identification of c-myc responsive genes using rat cDNA microarray. *Cancer Res* 2000; 60:5922-8.
8. Muller H, Bracken AP, Vernell R, Moroni MC, Christians F, Grassilli E, Prosperini E, Vignoli E, Oliner JD, Helin K. E2Fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis. *Genes Dev* 2001; 15:267-85.

9. Black EP, Huang E, Dressman H, Ishida S, West M, Nevins JR. Distinct gene expression phenotypes of cells lacking Rb and Rb family members. *Cancer Res* 2003; 63:3716-23.
10. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Jr, Marks JR, Nevins JR. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 2001; 98:11462-7.
11. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT. Gene expression predictors of breast cancer outcomes. *Lancet* 2003; 361:1590-6.