Genome-Wide Analysis of Genes, Gene Regulation and Gene Expression

Kathryn Brayer1, Olivia George1, Suzuki Hideaki1, Jason Byars1, Susan R. Atlas2, Jeremy Edwards3, Li Luo4, Huining Kang5, Christine Stidley6 and Scott Ness1,5

1Department of Internal Medicine, Section of Molecular Medicine
2Department of Physics & Astronomy, Center for Advanced Research Computing
3Department of Molecular Genetics and Microbiology
4Department of Internal Medicine, Division of Epidemiology
5UNM Cancer Center,
University of New Mexico, Albuquerque, NM 87131
E-mail: ssness@salud.unm.edu

Introduction – Alternative RNA Splicing

Human genes are discontinuous: they are composed of the protein-coding elements termed “exons” interspersed with non-coding portions called “introns”. During the process of gene expression, the exons and introns get transcribed into a precursor RNA but the introns then get removed through the process of “RNA splicing”, generating a final RNA product containing only the protein-coding exons. However, since many human genes contain dozens or even hundreds of exons, so each gene has the potential to produce multiple different RNA products if RNA splicing selectively skips or includes some exons.

Advances in genome-wide gene expression assays have shown that nearly all multi-exon genes undergo alternative RNA splicing and that the levels of alternative or aberrant RNA splicing are greatly increased in leukemia or tumor samples, compared to normal cells. Although some factors that regulate alternative splicing have oncogene-like transforming activity, it is not clear whether the increased levels of alternative splicing observed in tumors represent transformation-induced “noise” or whether alternatively spliced mRNAs encode variant proteins that actively contribute to the transformed phenotype.

For example, the human c-myb (MYB) gene (see figure at right) has 15 “normal” exons plus 6 additional “alternative” exons (shown in red in the figure) that get used only sometimes. When one or more of the alternative exons get used, the resulting RNAs encode different versions of the Myb protein, some of which are diagrammed at right. These variant proteins could have new properties or altered activities, and could play an important role in the development of leukemias. This same process occurs in many of the approximately 25,000 human genes, creating a highly complex mixture of variant proteins that contribute to important cell functions.

Assessing the importance of increased alternative RNA splicing on a genome-wide scale is problematic. The straightforward way to determine if the splicing events contribute to oncogenesis is to test the functions of the encoded proteins. However, the very large number of different transcripts that have been detected (>150,000) and the relatively low abundance of many of them complicates the analysis. How can one determine if the encoded variant proteins are important without expressing each one independently in normal cells or knocking them each out in tumors one-by-one? Furthermore, if the importance of the variant proteins is due to a mass action or combinatorial effect, then assessing their importance individually will not suffice.

Assessing the Importance of Alternative RNA Splicing

Our strategy to address this problem involves several elements: We are developing an innovative “deep” next-generation sequencing approach specially designed to determine the full-length structures of all the expressed transcripts. We are using this
technology to analyze a large, well-characterized cohort of high risk pediatric leukemia samples which we know contains multiple subgroups of patients with different outcomes and that will provide controls for our approach and a path to compare our results to other genomic methods. We will analyze both the “raw” transcript-level and exon-level data provided by the RNA sequencing, and we will resolve the transcript structures into predicted proteins to produce expression profiles for the transcripts, the exons and the predicted proteins. Finally, we will use advanced computational and statistical approaches to directly compare the importance of the RNA-based (e.g. transcript-level or exon-level) information to the protein-resolved data.

The next-generation sequencing will generate very large data sets (150-200 GB per sample) of raw sequence, which will require significant, customized processing to accomplish the goals of this project. Developing and implementing this data processing pipeline will be the responsibility of a team led by J Edwards, an engineer and molecular biologist who is an expert in the development of next-generation DNA sequencing technologies and SR Atlas, a physicist and computational scientist who is the Director of the UNM Center for Advanced Research Computing (the UNM supercomputer center), co-Director of the Biostatistics and Bioinformatics Shared Resource and who leads a team of programmers and database specialists expert in handling and analyzing large data sets. In addition, analysis of 150 samples will generate at least 20-40 TB of raw data. Unlike most genome sequencing projects, we will be writing custom algorithms to analyze the raw sequencing reads, so storing and handling this data will be an important function and will be the responsibility of Atlas and her team at the UNM Center for Advanced Research Computing (CARC).

As mentioned above, we will perform this analysis on at least 150 high risk childhood B-progenitor ALL samples from the Children’s Oncology Group P9906 cohort. Having access to the P9906 cohort samples provides unique opportunities and advantages. Since the same samples have been previously studied by gene expression profiling, we will be able to compare the results we obtain by analyzing alternative RNA splicing.

We found that resolving the alternative RNA splicing information into the predicted protein structures (e.g. counting all the transcripts that encode the same protein together) uncovered novel patient subgroups. For example, the figure at right shows that the predicted expression levels of the 95/10 variant of Myb divided these patients into two groups with distinct survival curves. These preliminary results suggest that this type of analysis, performed on a larger scale and with an informative cohort of samples, will successfully address the importance of increased alternative RNA splicing in leukemias. If the expression of specific proteins (based on resolving the RNA expression data into protein structures) correlates with outcome, the results would indicate that increased alternative splicing produces protein products that contribute to the transformed phenotype and that they are not merely due to “noisy” splicing.

By modifying the protocol for preparing RNA sequencing libraries, and by taking advantage of lower cost and higher throughput next-generation sequencing, we hope to determine whether alternative RNA splicing produces variant proteins that correlate with outcome, suggesting that they contribute to oncogenesis. These studies will provide the most detailed analysis of alternative RNA splicing in a clinically significant cohort of leukemia samples to date, and will produce valuable information about the processes and events leading to increased alternative RNA splicing in tumors. We are focusing our efforts on a cohort of high risk childhood B-progenitor ALL patients who routinely fail the standard therapies. Despite improvements in the treatments for B-ALL, this high risk cohort represents a group of patients for whom few good treatment options exist. The alternative splicing machinery contains many poorly characterized enzymes and regulatory proteins. If increased alternative splicing is found to play a role in tumor development, these proteins will represent novel and potentially “druggable” targets for the development of new therapies or interventions for these patients.

**Funding:**
NIH R01CA170250-01 Ness (PI) “Alternative RNA Splicing and Variant Protein Products in Leukemia Outcome (PQ11)”
NIH R01DE023222-01 Ness (PI) “Mutations and target genes in Adenoid Cystic Carcinoma”