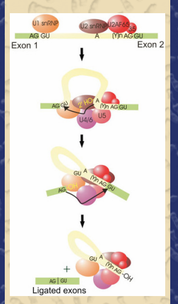
A microarray design for the detection of alternate isoforms
Application to a model of chemotherapy resistance in colon cancer

Malachi Griffith, Joseph Connors, Stephane Flibotte, Martin Hirst, Ryan Morin and Marco Marra

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

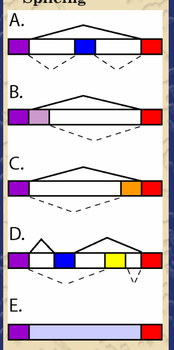
Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

Basic Splicing Machinery



RNA splicing is accomplished by the 'spliceosome' which consists of a complex of several nuclear components including a number of snRNPs. Through the action of RNA-RNA, RNA-protein and protein-protein interactions, this complex recognizes a pair of exon-intron boundaries and catalyzes sequential trans-esterification reactions to remove an intron and join two exons[1].

Types of Alternative Splicing



A.) Exon skipping/inclusion
B.) Alternate 5' splice site usage
C.) Alternate 3' splice site usage
D.) Mutually exclusive exons
E.) Intron retention

References

1.) Kalnina et al. 2005. Gen. Chrs. Cancer. 42(4): 342
2.) Johnson et al. 2003. Science. 302(5653): 2141-4
3.) Tai et al. 2005. J. Clin. Invest. 115(6): 1492-502
4.) Nuwaysir et al. 2003. Gen. Res. 12: 1749-55
5.) Pan et al. 2004. Mol. Cell. 16(6): 929-41
6.) Ule et al. 2005. Nat. Genet. 37(8): 844-52
7.) Castle et al. 2003. Genome Biol. 4(10): R66.
8.) Fehlbaum et al. 2005. NAR. 33(5): 37
9.) Le et al. 2004. NAR. 32(22): 180
10.)Andronescu et al. 2003. NAR. 31(13): 3416-22
11.)Gentleman et al. 2004. Gen. Biol. 5(10):R80.
12.)Bolstad et al. 2003. Bioinf. 19(2):185-93
13.)Bretz et al. 2005. Methods Inf. Med. 44(3): 431-7
14.)Tusher et al. 2001. PNAS. 98(9): 5116-21.
15.)Baross et al. 2004. Gen. Res. 14(10B): 2083-92

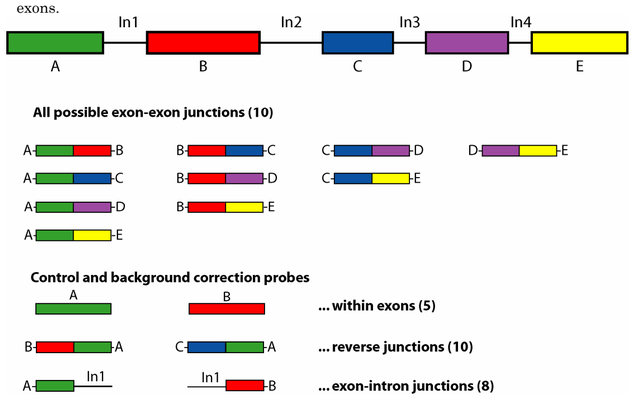
1. Abstract

The human genome contains approximately 25,000 genes. Traditional microarray studies of gene expression have generally operated under the simplifying assumption that each of these loci gives rise to a single protein isoform. Recent estimates suggest that as many as 74% of human genes undergo alternative splicing (AS). Developments in microarray technology now make it possible to comprehensively examine the expression of genes in the context of AS. Using a microarray approach similar to that described by Johnson et al. (2003) we aim to document the prevalence of AS and dissect the biological roles of proteins encoded by alternatively spliced transcripts[2]. We hypothesize that AS is an important mechanism for encoding a diversity of functions at a single genomic locus and that this diversity is realized in part through alterations in protein interactions. Furthermore, we hypothesize that changes in the pattern of AS are correlated with chemotherapy resistance. To address these hypotheses we propose to study AS in cell lines representing the transition from chemotherapy sensitivity to resistance in colon cancer[3]. This cell line model consists of a chemotherapy sensitive colon cancer cell line (MIP101) and clonally derived resistant cell lines established by long term exposure to four chemotherapy agents (5-FU, CPT-11, cisplatin, etoposide).

A number of recent experiments have used microarray technology to study AS. We recently designed arrays with 2.5 million oligonucleotide probes selected to measure exon combinations for ~17,250 human genes. A computational platform for AS Expression Analysis (ALEXA), consisting of Perl code and a mySQL relational database has been developed for the creation and analysis of AS microarrays. NimbleGen Systems Inc. will synthesize our arrays using their proprietary maskless photolithography procedure[4]. Our use of ~818,000 oligonucleotide probes for exon-exon junctions, along with ~231,000 exon-intron probes and ~155,000 exon-internal probes will enable identification of differentially expressed exon combinations with an unprecedented resolution and scale. With NimbleGen's current platform, six arrays will be required to accommodate probes for all protein coding genes. To identify exon combinations that are differentially expressed in chemotherapy resistant cells, raw signal intensities will be recorded in ALEXA, processed with Bioconductor, and analyzed using a modification of recent algorithms described for splicing microarrays[5,6]. Our objectives are to catalog exon combinations encoded by alternatively spliced transcripts specific to chemotherapy resistance and to characterize the differential expression, transcript structure and function of a set of transcripts with these exon combinations. Transcripts containing these novel combinations will be cloned, sequenced and analyzed. Transcript pairs will be prioritized according to their relevance to cancer and drug development and a subset of these used in proteomic studies to assess the extent to which protein interactions are altered by AS. One possible outcome to such work would be to use this catalog of AS events to generate a list of proteins for development of targeted therapies.

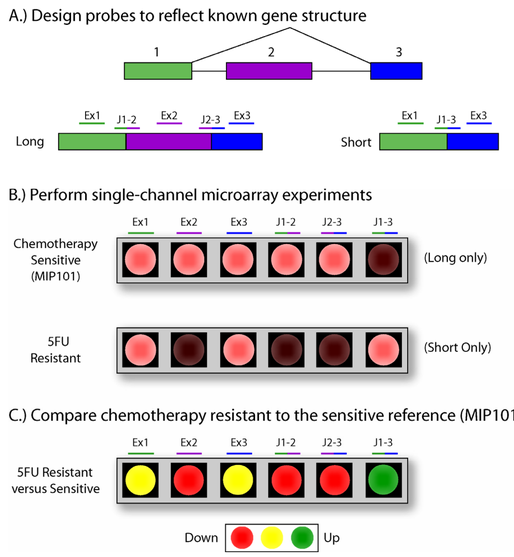
2. Array Design Strategy

We have created a computational platform for 'Alternative Expression Analysis' (ALEXA). ALEXA uses Perl and a mySQL relational database for the creation and analysis of AS microarrays. The database contains human genomic data from EnsEMBL, Entrez, UCSC, and oligonucleotide probe information. Currently, ALEXA contains data for 30,695 genes, 22,218 of which encode proteins. These protein coding genes are represented by 33,869 transcripts which use various combinations of the ~250,000 exons present in ALEXA. Of the 22,218 protein coding genes, 19,222 are annotated as having more than one exon and are therefore relevant to our AS study. The mean size of these transcripts is 2,500 bp and the mean number of exons per transcript is 10. The numbers of known transcripts per gene in ALEXA ranges from 1 to 12 but 68% of genes have only one. The remaining 32% (7,179) of the protein coding genes are represented by a total of 18,830 transcripts containing 13,315 exon-skipping events. The number of exons skipped in any particular event ranges from 1 to 142, but 95% of all exon skipping events observed involve a maximum of 5 exons. Until higher density array formats become available, our probes are limited to those that detect up to 10 skipped exons.



Our probe design was inspired by previous microarray studies of alternative splicing[5-9], but surveys a larger number of genes and transcripts, and places added emphasis on control probes. Probes are designed within every exon and across every possible exon-exon and exon-intron junction (illustrated above). Each exon junction probe is accompanied by a 'reverse-junction' probe which contains the same % GC sequence composition as the experimental probe but represents an invalid junction sequence. These will act as negative controls and allow an assessment of background hybridization. Probe design involved retrieval of genomic sequences from EnsEMBL, removal of single exon genes, pseudogenes, and non-coding RNA genes, masking of repeat elements using RepeatMasker and extraction of 36-mer probe sequences. The extracted probes were scored according to their specificity and predicted thermodynamic properties as follows: (1) melting temperature; (2) hairpin or dimerization potential as estimated by 'simFold' and 'pairFold' scores[10]; (3) presence of simple repeat elements identified by the mdust algorithm; (4) specificity of each probe using BLAST comparisons to all human mRNAs and EnsEMBL transcripts; and (5) specificity of each probe within the total population of probes. The following figure how our microarray design detects a simple splicing event.

Array Design Strategy Cont'd



3. Genome-Wide and Prototype Designs

Whole Genome Design

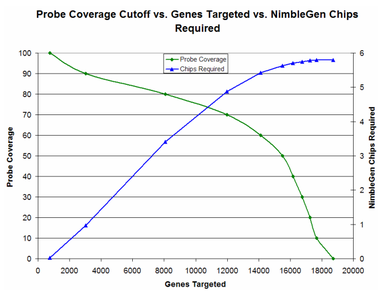
Of the 19,222 multi-exon protein-coding genes in ALEXA, 511 were excluded from the array design because they have more than 40 exons and would occupy an excessive amount of space on the microarray. The success of probe design varied considerably among the remaining 18,711 genes. To measure success, we considered the level of probe "coverage" as a ratio of successful probes to the ideal number needed to interrogate every possible splicing event for the known exons of each gene (refer to formula below). We identified 17,250 genes in which no gene had less than 20% coverage and where the average coverage was 75%. We selected 2.25 million probes for these genes, to be synthesized on six arrays. These probes target ~818,000 exon-exon junctions, ~231,000 exon-intron junctions and ~155,000 exons.

Test Array Design

A subset of 2,475 genes were selected to test our approach on a single NimbleGen array consisting of 385,000 probes. These genes were selected for their relevance to cancer, drug development efforts, or because they are known to be alternatively spliced in cancer. One third of these genes have known alternate transcripts according to EnsEMBL, one third have been identified as over-expressed in colon cancer by previous microarray studies and the average probe coverage for this subset of genes is 85%. Exon junction probes were limited to those that involve exon-skip events of 6 exons or less. This test design has been submitted to NimbleGen and is currently being synthesized.

Calculation of gene probe coverage, an estimate of probe design performance for a particular gene with **n** exons

$$\text{Probe Coverage} = \frac{\text{Successful Probe Count}}{\frac{\mathbf{n!}}{\mathbf{[n-2]!}} + [[\mathbf{n} \times 2] - 2] \times 2 + \mathbf{n}}$$

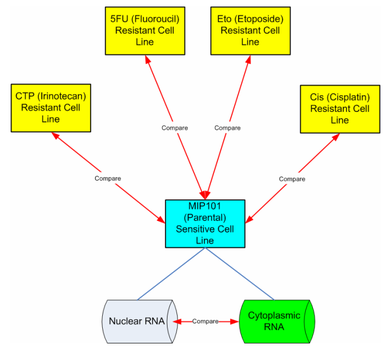


In order to cover the splicing events of most human protein-coding multi-exon genes (say ~17,500 of ~19,000 total), a minimum probe coverage of 20% is required. The resulting probes for these ~17,500 genes would occupy 6 NimbleGen arrays with 385,000 probes per array. Despite the low cutoff value, the average gene-probe coverage is still 75%.

4. Sample preparation and hybridizations

Using the approach described here we plan to study the role of alternative splicing in a colorectal cancer cell line representing the transition from chemotherapy sensitivity to resistance[3]. Chemotherapy resistant colorectal cancer cell lines were established from a sensitive parental line (MIP101) by long term exposure to four chemotherapy agents (5-FU, CPT-11, cisplatin and etoposide). These clonally derived cell lines exhibit at least an 8-fold increase in IC50 compared to the parental MIP101 cell line. The clonal derivation of these lines is a major advantage, eliminating differences in drug sensitivity due to fundamental genetic differences between the lines. The following figure depicts the comparisons of interest for this cell line model. Cytoplasmic RNA will be isolated from the sensitive parental line and each of the four resistant derivatives using Ambion's PARIS kit (Cat #1921). Total cytoplasmic RNA will be DNAse treated and used for isolation of mRNA. Approximately 10 µg of mRNA will sent to NimbleGen for labeling and hybridization experiments. Labeling will be conducted using a random priming approach to label the whole transcript (thus avoiding the 3' bias often associated with oligo-dT labeling). NimbleGen will conduct hybridizations and send us unprocessed signal intensity values.

5. Experimental Design



Hybridization experiments will be conducted in triplicate. RNA isolated from cytoplasmic and nuclear fractions will be compared to estimate the potential presence of 'splicing noise' caused by immature transcripts from the nucleus. Pairwise comparisons between each chemotherapy resistant cell line and the sensitive parental cell line will be conducted to identify differentially expressed exon combinations specific to chemotherapy resistance

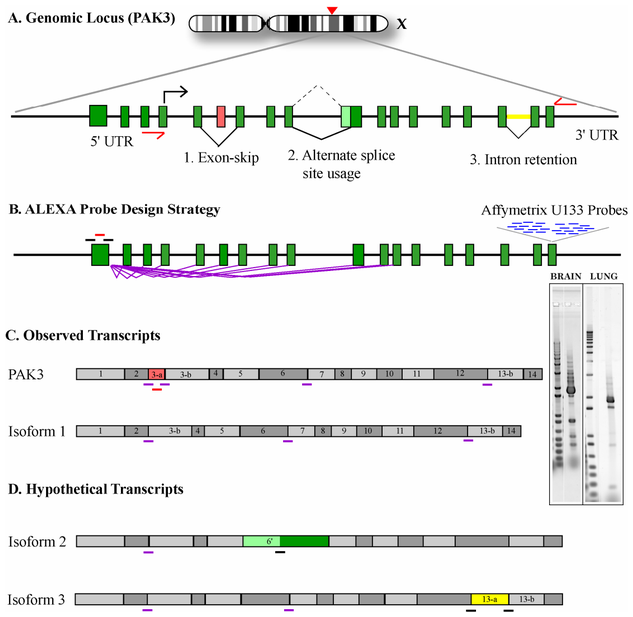
6. Analysis and Validation

Background Correction, Normalization and Differential Expression Analysis

We will record into ALEXA the raw signal intensities measured for all control and experimental oligonucleotides. Background signal intensity levels will be estimated by analysis of the reverse-junction oligonucleotide probes, one of which is present for each exon junction probe. Raw intensity data from single channel hybridization experiments produced by NimbleGen will be processed primarily with Bioconductor packages written in the statistical programming language, R[11]. Global and probe specific background intensity levels will be estimated from reverse-junction and exon-intron junction probes for which no true hybridization signal is expected. Quantiles normalization will be used for probe level normalization of all single channel data sets[12]. Differential expression can be assessed at the level of exons and exon pairs by modifying methods such as Bioconductor's 'siggenes' and 'multtest' packages[13,14]. Because our array design contains a unique combination of features, novel modifications of standard procedures for background correction, normalization, and identification of differential expression may be required.

Validations

Differentially expressed exon combinations predicted from microarray data will be validated by quantitative real-time PCR. We will design primers to generate short amplicons spanning exon-exon junctions and these will be combined in quantitative RT-PCR reactions with RNA and a non-specific fluorescent dye. A separate validation will involve parsing the millions of sequences in publicly available transcript sequence databases (ESTs, RefSeq, MGC, etc.) for exon combinations detected by the arrays. This will allow us to determine whether differentially expressed exon combinations were detected before, and if so, in what tissue. Splicing events that are successfully validated by this approach will be prioritized for recovery of full open reading frame clones[15] representing transcripts containing the differentially expressed exon combination (see example below). We aim to clone at least two transcripts for each differentially expressed exon combination, one containing the differentially expressed exon combination and the other will lacking that particular combination. Cloning targets will be prioritized by their statistical significance of differential expression, known relevance of the genes to cancer, and extent to which the proteins may be "druggable". Recovered clone pairs will be provided to the GSC proteomics group for a high-throughput comparison of their protein-protein interactions.



7. Acknowledgments

In addition to the authors listed above I would like to thank Obi Griffith, Steven Jones, Diana Mah, Michelle Moksa, Gregg Morin, Johnson Pang, Anna-Liisa Prabhu, Adrian Quayle, Marianne Sadar, Isabella Tai, Michelle Tang, Kane Tse, Jing Wang, Thomas Zeng and Yongjun Zhao

Michael Smith Foundation for Health Research
NSERC CRSNG