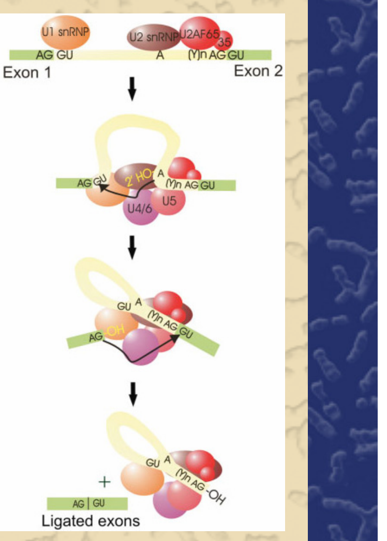
Characterization of Alternative Splicing Events in the Druggable Genome Using Splicing Microarrays

Malachi Griffith, Stephane Flibotte, Thomas Zeng, Tammy Romanuik, Marianne Sadar, Martin Hirst and Marco Marra

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

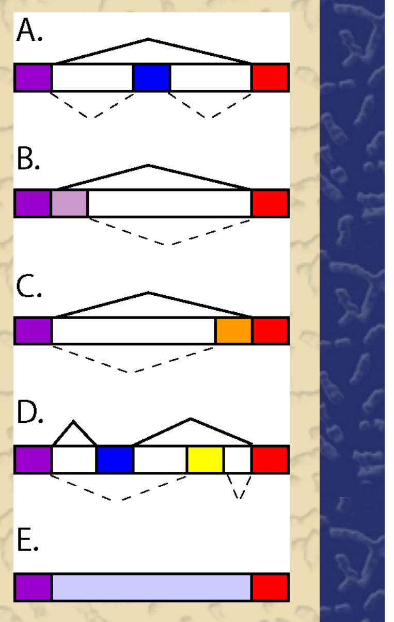
Canada's Michael Smith
Genome Sciences Centre
www.bcgsc.ca

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 3' splice site usage
C.) Alternate 5' splice site usage
D.) Mutually exclusive exons
E.) Intron retention

1. Abstract

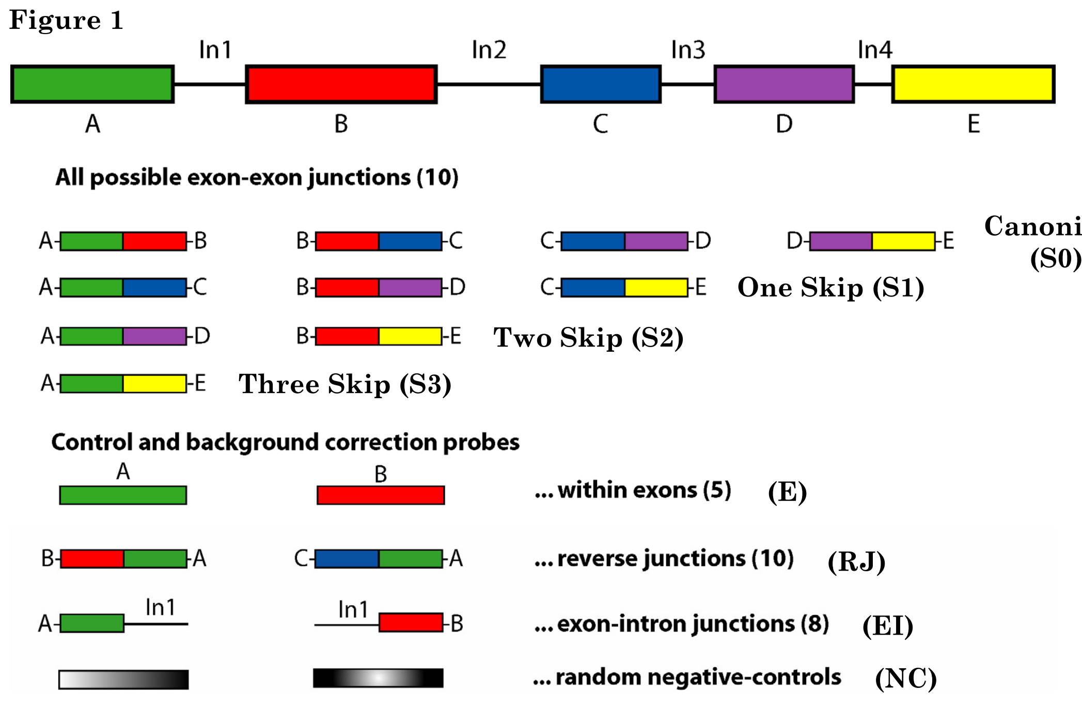
Introduction: The human genome contains approximately 25,000 genes. Traditional microarray studies of gene expression have operated under the assumption that each of these loci generally gives rise to a single isoform. Recent estimates suggest that as many as 74% of human genes undergo alternative splicing (AS)[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-protein interactions or sub-cellular location.
Methods: We have designed microarrays consisting of 2.5 million 36-mer oligonucleotide probes selected to measure potential AS events for ~17,250 multi-exon, protein-coding human genes. A computational platform for **AS E**xpression **A**nalysis (ALEXA) has been developed for the creation and analysis of such microarrays for any species in the EnsEMBL database. The complete human design consists of ~818,000 oligonucleotide probes for exon-exon junctions, along with ~231,000 exon-intron probes and ~155,000 exon-internal probes. This level of probe coverage will enable identification of differentially expressed exon combinations with a resolution and scale exceeding that of current microarray expression platforms. Additional elements of the design include an emphasis on control probes such as the creation of a 'reverse-junction' probe for every experimental probe and a set of randomly generated probes selected to uniformly represent the thermodynamic parameters of experimental probes. To test the effectiveness of this design, preliminary experiments were conducted using a prototype array synthesized by NimbleGen Systems Inc[3]. This array consisted of 385,000 probes corresponding to a subset of ~2,500 human genes representing potential therapeutic drug targets. Preliminary experiments to test the sensitivity and specificity of the custom array involved hybridizations of randomly labeled mRNA isolated from normal brain tissue (cerebellum) and a prostate cancer cell line (LnCAP).
Results: Hybridization data was analyzed by recording raw signal intensities in the ALEXA database and processing with Bioconductor[4] and novel algorithms. Preliminary analysis has focused on: (1) evaluating the behaviour of experimental and control probe types (2) comparing alternate hybridization conditions (3) assessing the effectiveness of background correction strategies and (4) identifying novel and known AS events with evidence for expression in a single condition or differential expression between conditions.
Conclusions: Our objectives are to catalog the prevalence and diversity of AS events among genes of the druggable genome[5]. To characterize the differential expression, transcript structure and function of a subset of these events, transcript pairs representing AS events predicted from the microarray data will be cloned and sequenced. The approach described here represents an extension of current microarray design strategies with the ability to identify novel alternatively spliced isoforms representing potential novel therapeutic drug targets.

2. Array design & preliminary experimentation

Array design

We designed probes within every human exon and across every possible exon-exon and exon-intron junction (**Figure 1**). 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 probes act as an estimate of false positive predictions. A set of random probe sequences were also generated to uniformly cover the Tm range of experimental probes to 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 and filtered to optimize the following parameters: melting temperature (Tm); hairpin or dimerization potential; presence of simple repeat elements; specificity within all human mRNAs and EnsEMBL transcripts and specificity of each probe within the total population of probes.

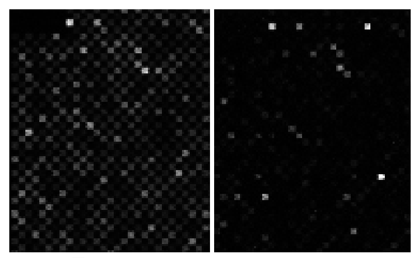
Figure 1


In1
In2
In3
In4
A
B
C
D
E
All possible exon-exon junctions (10)
A-B
B-C
C-D
D-E
Canonical (S0)
A-C
B-D
C-E
One Skip (S1)
A-D
B-E
Two Skip (S2)
A-E
Three Skip (S3)
Control and background correction probes
A
B
... within exons (5)
(E)
B-A
C-A
... reverse junctions (10)
(RJ)
A-In1
In1-B
... exon-intron junctions (8)
(EI)
... random negative-controls
(NC)

Pilot hybridizations with a prototype array

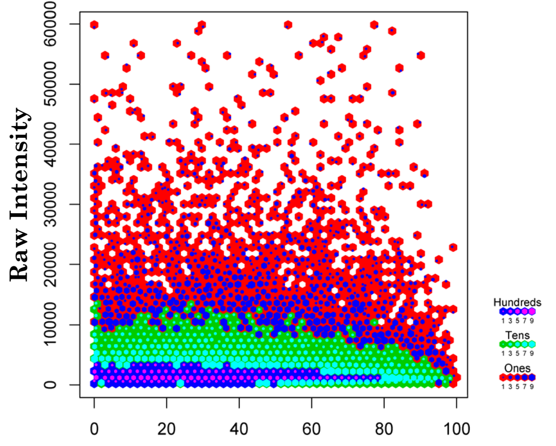
A design file containing 385,000 probe sequences for ~2,500 "druggable" genes was submitted to NimbleGen for layout and synthesis. mRNA isolated from normal brain tissue (cerebellum) and prostate cancer cell lines (LnCAP) was sent to NimbleGen and labeled using a random hexamer approach. Each sample was hybridized using one of two hybridization conditions: (1) optimized for 24-mer probes (low stringency); (2) optimized for 60-mer probes (high stringency). A plot of the raw intensity values for all exon probes for genes larger than 3000 bp shows no significant relationship between the position of a probe and observed intensity (**Figure 3**). Unless otherwise indicated, plots show high stringency data for the LnCAP mRNA sample.

Figure 2



Low Stringency Hybe
High Stringency Hybe

Figure 3



3. Data analysis

Normalization and preliminary analysis of probe behaviour

Hybridization data was normalized by quantiles normalization[6] between samples hybridized at the same stringency. A pool of ~4,400 negative-control probes with random sequences were used to calculate the background hybridization level for experimental probes. **Figures 4** and **5** show the distribution of all exon (**blue**) and negative-control (**orange**) probe log2 intensities for the low and high stringency hybridization conditions respectively.

Figure 4 – Low Stringency

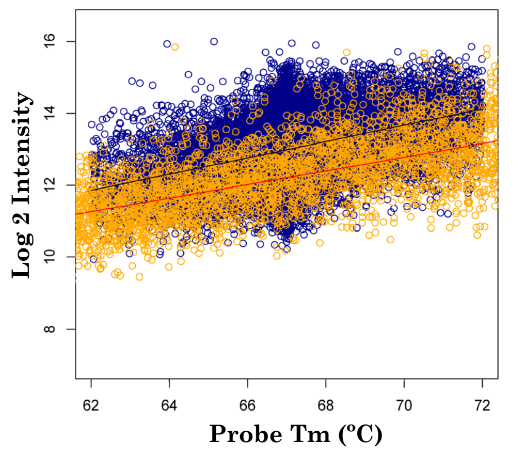


Figure 5 – High Stringency

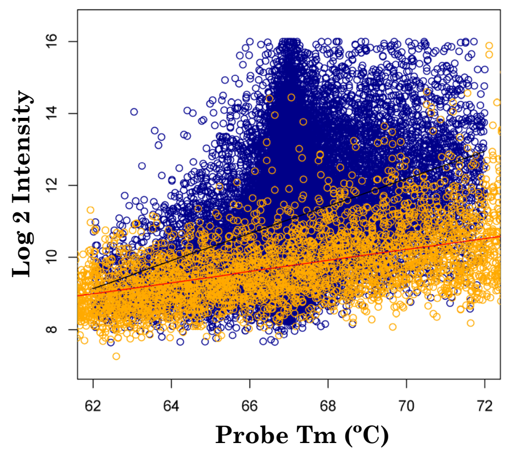
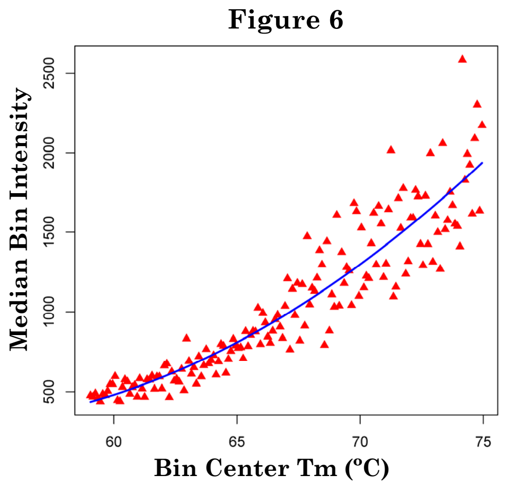


Figure 6



Background correction

Negative-control probes were divided into bins according to their Tm (range of 0.1 °C per bin), filtered to remove outliers and the median intensity of the remaining probes plotted against Tm (see **Figure 6**). A polynomial quadratic model was fit to these data for each hybridization experiment and the resulting model fit was used with the Tm of each experimental probe to estimate the expected level of background hybridization intensity (**blue** line shows model fit).

Quadratic model

$$I = (a \times Tm^2) + (b \times Tm) + c$$

Differential expression

Background corrected (BGC) values were calculated as a ratio of the normalized probe intensity to the background intensity estimate. Probe-level differential expression (DE) values were calculated as a log2 ratio of the BGC values between the two samples. Gene-level DE values were calculated as the median of all exon and canonical probe DE values.

4. Results

Increased hybridization stringency yields improved specificity

The following panel shows a comparison between the observed intensity values for a highly expressed gene (ATP Citrate Lyase) in the LnCAP sample. Intensity values are color coded according to probe type (refer to legend). **Figures 7** and **8** show the low and high stringency hybridization data respectively. Dotted lines indicate: (mean of RJ probe intensities + (2 × SD))

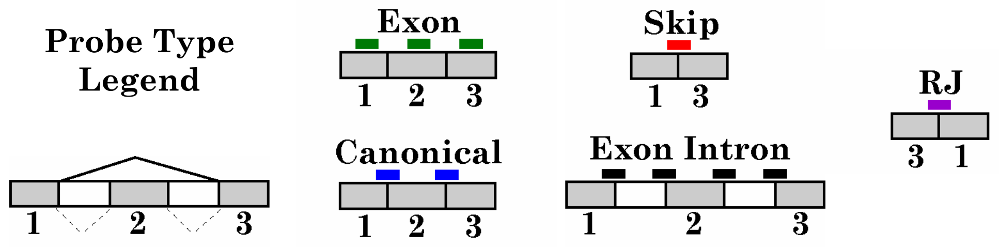


Figure 7 – Low Stringency

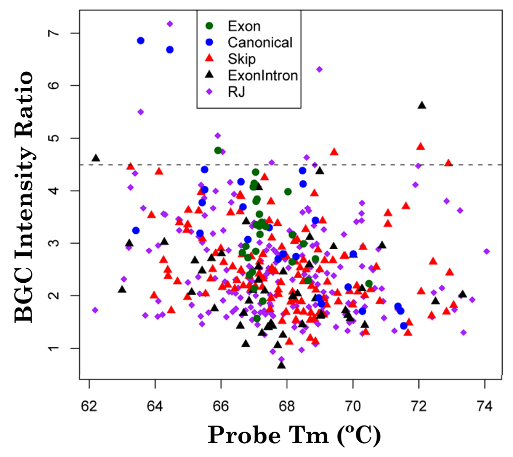
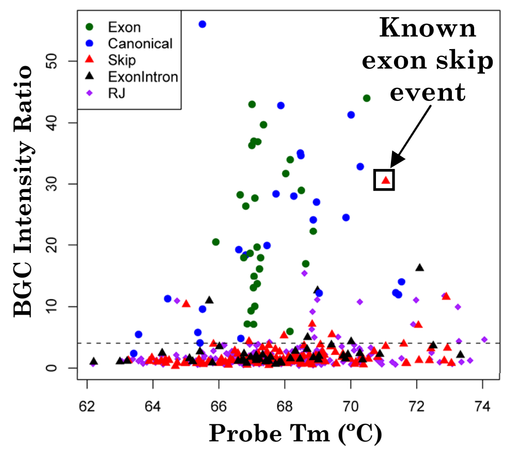


Figure 8 – High Stringency



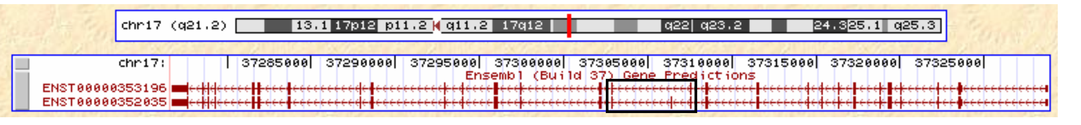


Figure 9 – High Stringency

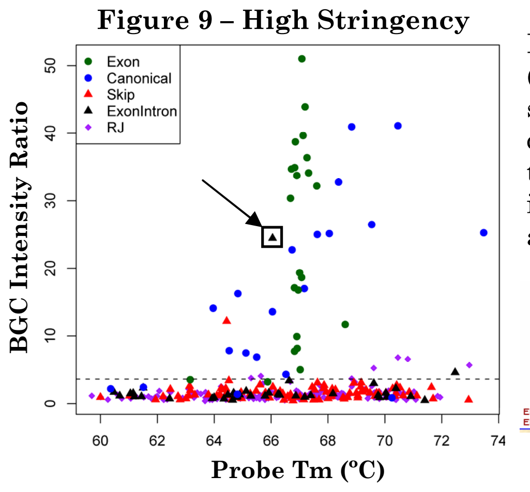
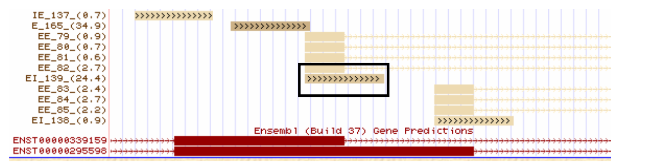


Figure 9 shows a second example of a gene (ATP1A1) which is highly expressed in the LnCAP sample. Note the difference between intensities observed for canonical and exon probes versus those representing putative AS events. The exon-intron probe indicated corresponds to a known alternate exon boundary (shown below).



4. Results (cont'd)

Profiling the exon boundaries of the Prostate Specific Antigen (PSA) gene

DE values for a number of exon and canonical probes indicate high expression of the PSA gene in LnCAP relative to brain (**Figure 10**). A number of exon-intron probes also have high DE values, potentially indicating alternate splice site usage events. Existing transcript data (**Figure 11**) and recent publications[7,8] also suggest an unusually complex pattern of splicing for this gene. **Figure 11** also shows (in **green**) the position of 454 sequencing reads we generated using the same LnCAP RNA profiled on the arrays[9]. Of ~180,000 reads generated for the LnCAP sample, 388 mapped to the PSA locus and a number of these support the novel and known alternate splice site usage events predicted by the array data.

Figure 10 – High Stringency

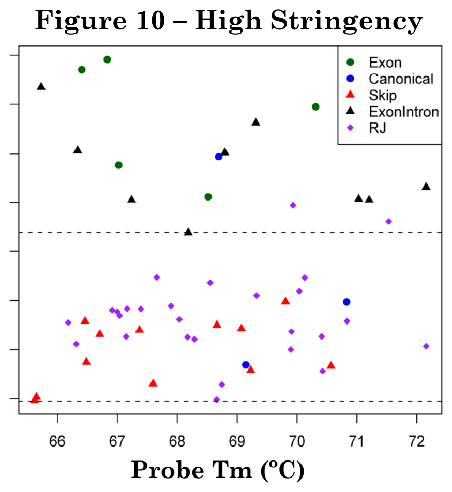


Figure 11

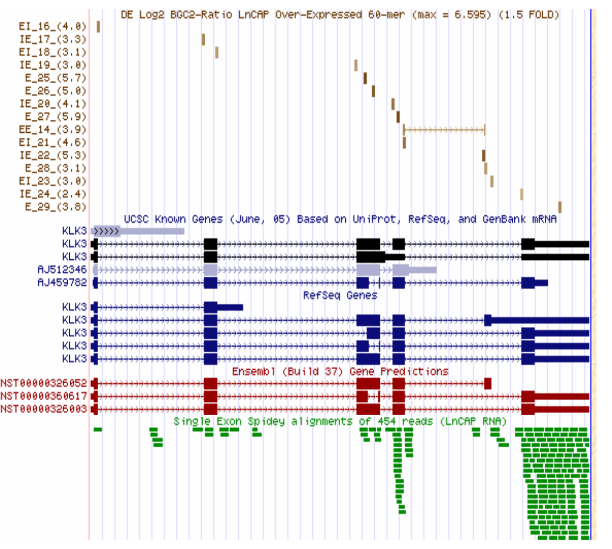
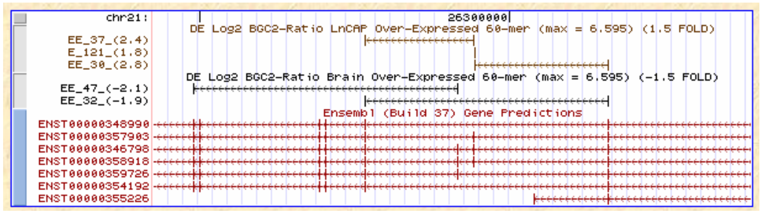


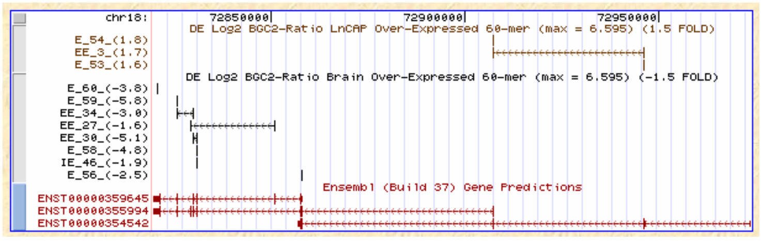
Figure 12


chr21:
26300000
DE Log2 BGC2-Ratio LnCAP Over-Expressed 60-mer (max = 6.595) (1.5 FOLD)
EE_37_(2.4)
E_121_(1.8)
EE_30_(2.8)
DE Log2 BGC2-Ratio Brain Over-Expressed 60-mer (max = 6.595) (-1.5 FOLD)
EE_47_(-2.1)
EE_32_(-1.9)
Ensembl (Build 37) Gene Predictions
ENST00000348990
ENST00000357903
ENST00000346798
ENST00000358918
ENST00000359726
ENST00000354192
ENST00000355226

Differentially expressed AS events in Amyloid beta Precursor Protein (APP)

Figure 12 shows the position of multiple probes with evidence for DE of the same AS event.

Figure 13



Differential expression of Myelin Basic Protein (MBP) isoforms

8 probes indicate over-expression of a short MBP isoform in Brain and 3 probes indicate over-expression of a long isoform in LnCAP.

Table 1 – Percentage of probes with DE values > 2-fold (by probe type)

	Low	High
E	1.64%	25.62%
S0	2.65%	21.98%
S1	1.06%	4.38%
S2	1.11%	3.81%
S3	0.93%	3.82%
S4	0.92%	3.88%
S5	0.85%	3.72%
S6	0.84%	3.94%
EI	0.63%	4.58%
NC	0.17%	2.35%

Table 1 shows the percentage of probes with evidence for DE > 2-fold for all exon probes (E), canonical probes (S0), exon-skipping probes (S1-S6), exon-intron probes (EI) and negative-control probes (NC). The percentage of NC probes gives an estimate of the false positive rate in other probe categories. Data for both the low and high stringency hybridizations are shown.

Algorithms for identifying differentially expressed AS events by combining intensity values from multiple probes that indicate the inclusion or exclusion of an exon-combination are currently under development. The utility of these algorithms will be assessed by comparing the number of known AS events detected relative to the total number of AS events predicted. The prototype array described in this work contains probes for ~1,100 known exon skipping junctions, ~1,100 known alternate exon-intron boundaries, and ~4,500 non-constitutive exons.

5. Conclusions and future work

A subset of the novel and known differentially expressed AS events observed in the array data described above will be selected for validation by qRT-PCR. Although the microarray analysis we have developed is capable of revealing differentially expressed exons and exon junctions, it does not reveal the structure of the complete transcripts bearing these differentially expressed features. To reveal complete transcript structures, we will generate sets of 12 clones for each of a subset of "druggable" genes identified as bearing a differentially expressed feature. To generate the clones we will design primers flanking the longest known ORF for each gene, and use these to generate amplicons from our RNA samples. These full-ORF amplicons will be cloned, sequenced and annotated by a high-throughput pipeline currently in operation at the Genome Sciences Centre[10].

6. Acknowledgments

Thanks | In addition to the authors listed above I would like to thank the GSC Mammalian Gene Collection Closure Group, Obi Griffith, Anca Petrescu, Michelle Tang, Isabella Tai, Steve Jones, Joseph Connors, and Gregg Morin

Funding | I am grateful for support provided by the Natural Sciences and Engineering Research Council, the Michael Smith Foundation for Health Research and Genome British Columbia

NSERC CRSNG
Michael Smith Foundation for Health Research
GenomeBritishColumbia

References

1.) Kalnina et al. 2005. Gen. Chrs. Cancer. 42(4): 342
2.) Johnson et al. 2003. Science. 302(5653): 2141-4
3.) Nuwaysir et al. 2003. Gen. Res. 12: 1749-55
4.) Gentleman et al. 2004. Gen. Biol. 5(10):R80.
5.) Hopkins et al. 2002. Nat Rev Drug Discov. 1(9): 727-30
6.) Bolstad et al. 2003. Bioinf. 19(2):185-93
7.) Heuze-Vourch et al. 2003. Eur. J. Biochem. 270: 706-14
8.) David et al. 2002. J. Bio. Chem. 277(20): 18084-90
9.) Bainbridge et al. 2006 Submitted BMC Genomics
10.) Baross et al. 2004. Gen. Res. 14(10B): 2083-92