BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

SplicerAV: a tool for mining microarray expression data for changes in RNA processing

Timothy J Robinson[1], Michaela A Dinan[2], Mark Dewhirst[3], Mariano A Garcia-Blanco[4,5,6*], James L Pearson[4,6*]

Abstract

Background: Over the past two decades more than fifty thousand unique clinical and biological samples have been assayed using the Affymetrix HG-U133 and HG-U95 GeneChip microarray platforms. This substantial repository has been used extensively to characterize changes in gene expression between biological samples, but has not been previously mined *en masse* for changes in mRNA processing. We explored the possibility of using HG-U133 microarray data to identify changes in alternative mRNA processing in several available archival datasets.

Results: Data from these and other gene expression microarrays can now be mined for changes in transcript isoform abundance using a program described here, SplicerAV. Using *in vivo* and *in vitro* breast cancer microarray datasets, SplicerAV was able to perform both gene and isoform specific expression profiling within the same microarray dataset. Our reanalysis of Affymetrix U133 plus 2.0 data generated by *in vitro* over-expression of HRAS, E2F3, beta-catenin (CTNNB1), SRC, and MYC identified several hundred oncogene-induced mRNA isoform changes, one of which recognized a previously unknown mechanism of *EGFR* family activation. Using clinical data, SplicerAV predicted 241 isoform changes between low and high grade breast tumors; with changes enriched among genes coding for guanyl-nucleotide exchange factors, metalloprotease inhibitors, and mRNA processing factors. Isoform changes in 15 genes were associated with aggressive cancer across the three breast cancer datasets.

Conclusions: Using SplicerAV, we identified several hundred previously uncharacterized isoform changes induced by *in vitro* oncogene over-expression and revealed a previously unknown mechanism of EGFR activation in human mammary epithelial cells. We analyzed Affymetrix GeneChip data from over 400 human breast tumors in three independent studies, making this the largest clinical dataset analyzed for *en masse* changes in alternative mRNA processing. The capacity to detect RNA isoform changes in archival microarray data using SplicerAV allowed us to carry out the first analysis of isoform specific mRNA changes directly associated with cancer survival.

Background

The key postulate that one gene encodes one polypeptide chain (one enzyme) has been overhauled with the discovery that one gene can generate multiple RNA transcripts (and indirectly many different polypeptide chains) through a process referred to as alternative mRNA processing [1]. Alternative processing defines a range of events, including alternative splicing and alternative polyadenylation, which result in distinct mRNA species. Recent deep sequencing studies indicate that 94% of all protein coding genes generate multiple mRNA transcripts [2] and mutations affecting mRNA splicing are responsible for an estimated 15-60% of human genetic diseases [3,4]. Functional consequences of alternative processing have been shown across a wide variety of biological processes (reviewed by [5-7]) including drug metabolism, stem cell renewal, neurologic disease, autoimmune disease, and especially cancer. Despite the importance of alternative processing in cancer, current understanding of its global regulation remains sparse [8] and limits the ability to fully harness alternative processing as a tool in cancer prognosis, diagnosis, and treatment.

Attempts to obtain a genome scale understanding of alternative processing in cancer have focused on large-scale characterizations of changes in alternative processing between normal tissue and cancer. Bioinformatic analyses have identified a large number of transcript isoforms found only within cancer tissue [9-11]. The recent

* Correspondence: garci001@mc.duke.edu; james.pearson@duke.edu
[4]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, USA

use of splicing sensitive microarrays has allowed quantification of changes in alternative processing between individual samples (reviewed in [1]). These arrays have been used to detect changes in alternative processing between normal human tissues and in breast, brain, colon, prostate, and bladder carcinomas [12-16] using various splicing algorithms (reviewed in [17]). Large scale clinical analyses of changes in alternative processing; however, remain sparse, and there are no high-throughput analyses of changes in mRNA processing associated with poor patient prognosis. Such studies require years of patient follow-up and have not been reported using the new splicing arrays.

In contrast, public repositories such as the Gene Expression Omnibus (GEO) currently contain conventional gene expression data from hundreds of thousands of unique biological or clinical samples ([18]). Data previously generated by the microarray community provide an untapped source of potential insight to the regulation of alternative mRNA processing in human cancer. Although the exact value of these data is not known, it is likely that well over a billion dollars have been invested in reagents, facility, and personnel costs over the past two decades.

The first commercially available high-density gene expression microarrays were invented three decades ago by Affymetrix [19] to quantify expression changes in tens of thousands of genes in a single experiment, but were not intended to detect isoform specific mRNA changes resulting from alternative processing. Two of the most commonly used human expression microarrays, the Affymetrix U95 and U133 series, use individual probesets to report expression of many genes. Each probeset is composed of 11 individual 25 nt oligomers that interrogate a subsequence of the target gene. Both platforms, however, contain thousands of genes whose expression is assayed by more than one probeset. The use of multiple probesets, which often interrogate non-overlapping regions of the target gene, was originally intended to provide a robust assay of gene expression. We and others have previously observed that discrepancies between fold-changes in probesets interrogating the same gene can represent isoform-specific changes in mRNA levels [20-22]. Such isoform changes can result from alternative transcription start sites, alternative mRNA processing, or changes in mRNA isoform stability.

Methods that detect isoform-specific mRNA changes have been developed for splicing microarrays such as the Affymetrix Human Exon 1.0 ST (reviewed in [17]), but have not been developed for or applied to conventional gene expression microarrays. In fact, it has been suggested in such reviews that "detection of disease-relevant splicing differences may be entirely missed in gene-level expression profiling studies" [17]. Although it may be possible in theory to apply such methods to conventional gene expression microarrays, to our knowledge this has not been done. To fully investigate the potential to detect isoform-specific mRNA changes in conventional gene expression microarray data, we elected to develop a novel method, SplicerAV, which we have applied to conventional Affymetrix gene expression microarray data.

For the Affymetrix GeneChip Human U133 plus 2.0 arrays, 11,193 genes, which represent 57% of uniquely annotated genes assayed by the array, are interrogated by multiple probesets and can therefore be queried for mRNA isoform changes, with an average of 3.2 probesets interrogating these genes (Table 1). For the U133A arrays, 36% are interrogated by multiple probesets, with an average of 2.7 probesets per gene for a total of 4,609 genes. The U133 series of array platforms are among the most commonly used platforms within GEO (over 40,000 samples) and have the potential to detect isoform changes in thousands of genes.

SplicerAV is a program created to systematically assess the likelihood of changes in alternative processing evidenced by discrepancies in probeset behavior using a Gaussian mixture model of mRNA transcript regulation. A beta version of this program, which lacked biological modifiers and the ability to generate estimates of statistical significance, was initially used to identify differential regulation of transcript isoforms by *TCERG1* [20]. SplicerAV can be applied to any expression microarray platform with multiple probesets interrogating the same gene, without the need for detailed transcript annotation. The program provides a non-computationally intensive algorithm capable of analyzing probeset-summary level datasets for evidence of changes in alternative mRNA processing. We provide here a description of SplicerAV, which has been developed to provide a rigorous statistical model and incorporate biologically motivated modifications with the goal of assisting biologists in identifying alternative processing events most amenable for in-depth study from conventional gene expression microarray data.

In this study SplicerAV's unique value in detecting previously overlooked changes in mRNA processing is demonstrated using publicly available Affymetrix U133 gene expression datasets. SplicerAV was used to uncover previously uncharacterized isoform specific changes in epidermal growth factor receptor (*EGFR*) caused by *in vitro* HRAS over-expression [23]. In a separate analysis, SplicerAV was used to identify changes in alternative mRNA processing associated with poor patient prognosis in over 400 breast tumors. Here we demonstrate SplicerAV's ability to examine archival data, performing the largest analysis of alternative mRNA processing in

Table 1 SplicerAV related probeset features of commonly used Affymetrix microarrays

Platform	Unique Annotated Genes	Genes w/Mult Probesets	Fraction of genes w/mult probesets	Avg. Probesets per gene	Unannotated Probesets	Total Probesets
U133 Plus 2.0	19,761	11,193	57%	3.2	9818	54,675
U133 A	12,737	4,609	36%	2.7	1917	22,283
U95 A	8,690	1,946	22%	2.4	1253	12,651
Mouse 430A 2	12,755	4,934	39%	2.6	2118	22,690

human cancer to date and the only high-throughput analysis of changes in alternative mRNA processing associated with human cancer prognosis.

Results and Discussion

SplicerAV Algorithm

There are two main steps in the SplicerAV analysis. The first step summarizes individual probeset changes in expression between a user defined group of control and treatment observations. The second step evaluates these probeset level summaries for evidence of changes in alternative processing using a Gaussian mixture model (Figure 1).

In the first step, changes in probeset expression levels are summarized by calculating their average $\log_2$fold changes and corresponding t-statistics. These metrics were taken from conventional gene expression analysis. Probesets targeting the same gene are then grouped together and each probeset is assigned a weight. Individual probeset weights are calculated using a combination of that probeset's t-statistic, number of observations, and comparison with other probesets targeting the same gene (see methods).

Once these weights are assigned, each gene is evaluated for evidence of alternative processing using a Gaussian mixture model. In the Gaussian mixture model used by SplicerAV, probesets interrogating a transcriptionally activated gene are predicted to detect the same proportional increase in expression. For example, probesets targeting an mRNA that doubles in abundance would be expected to double in intensity (Figure 1B). Conversely, probesets targeting an mRNA which is down-regulated by half would be expected to be reduced by half Figure 1C). Multiple probesets targeting a gene that is alternatively processed or undergoes isoform specific mRNA regulation would be expected to report discordant changes in probeset intensities (Figure 1D).

Plotting the same aforementioned hypothetical data as $\log_2$ fold-changes emphasizes that in alternatively processed mRNAs, summarized probeset behavior clusters into discrete groups (Figure 1, right). SplicerAV assesses this grouping mathematically assuming a Gaussian mixture model, which compares fitting the data using one vs. two Gaussian distributions. Fitting the probeset expression data with a single Gaussian curve equates to a biological model in which the gene is regulated as one expression unit (e.g., all transcripts are destabilized equally). Fitting the data with a two Gaussian model equates to a biological model in which the gene is regulated as two or more expression units, corresponding to changes in isoform specific regulation. Comparing the ratio of how well each model fits the summarized probeset data gives a maximum likelihood ratio, or MLR, which gives an indication of how well the summarized

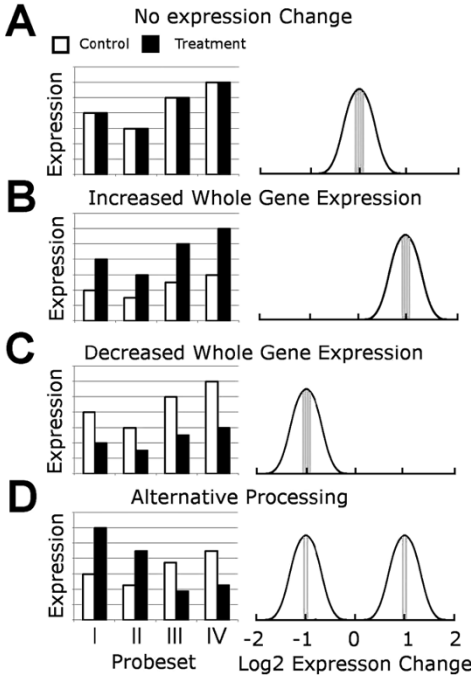

Figure 1 Gaussian mixture model of changes in alternative processing. Absolute expression of a hypothetical gene is reported by four independent probesets targeting different regions of this gene; I, II, III, IV (left panels) for control and treatment conditions (open and closed bars respectively). The idealized Gaussian mixture models representing changes in probeset behavior are illustrated in the right panels. Panels A, B, and C represent concordant probeset behaviors corresponding to no change, an increase, and a decrease, respectively. Panel D represents discordant behavior; two probesets (I, II) report an increase, while the remaining probesets (III, IV) report a decrease in expression between conditions (control and treatment). Probesets may report discrepant changes in gene expression depending on which region of the mRNA transcript they interrogate.

probeset data are described by changes in alternative processing relative to whole transcript regulation. The lowest possible log MLR for a gene is zero, which indicates that all probesets change proportionally and suggests no evidence of alternative processing. Log MLRs greater than zero indicate discrepancy in the expression changes in the probesets, which can be caused by an alternative processing event. The greater the value of the log MLR the more likely a gene is to be alternatively processed (see methods for more details).

$$MLR = \frac{(Likelihood\ of\ probeset\ data|Two\ Gaussian\ Curves)}{(Likelihood\ of\ probeset\ data|Single\ Gaussian\ Curve)} \quad (1)$$

SplicerAV uses the chip annotation file ("platform_annot.csv" for Affymetrix arrays) to determine which probesets interrogate the same gene. For most microarray platforms the gene symbol provides an appropriate annotation scheme, however any provided annotation (Transcript cluster ID, WormBase, FlyBase, Ensembl, etc.) can be used.

Probeset Annotation & Filtering

Our analyses used the default probeset annotation provided by Affymetrix. This annotation contains probesets that in some cases target multiple exons or are poorly annotated [24-26]. Re-defining probeset definition, for example using exon-based definitions of probesets, may improve the ability of SplicerAV to detect changes in mRNA processing [24,25]. However, using the standard annotation provided by Affymetrix makes our findings here directly comparable to the vast majority of expression analyses conducted using the U133 series of arrays, allowing reference to specific probeset IDs and enabling us to directly analyze summarized expression datasets deposited in GEO. Additionally, many Affymetrix microarray expression datasets deposited in GEO do not contain CEL files [26] and cannot be re-analyzed using custom annotation.

The use of standard Affymetrix annotation also allows us to make presence/absence probeset detection calls using previously validated methods [27]. As described above, SplicerAV detects discrepancies in fold changes between probesets targeting the same gene, using these discrepancies to infer changes in alternative mRNA processing. Nevertheless, such discrepancies can also reflect the presence of negative strand matching probesets (NSMPs) or probesets that do not produce signal above background, which can be caused by low transcript levels or non-functional probes. NSMPs hybridize or detect RNAs transcribed in the opposite direction of the annotated gene; they do not reflect the expression of the target transcript and are identified and removed by SplicerAV using information available in standard Affymetrix annotation files [27]. Probesets that do not produce signal can also falsely suggest isoform specific mRNA changes. These probesets are removed by SplicerAV if they are not expressed above background ($P < .05$) in either treatment or control groups using the Presence-Absence calls with Negative Probesets (PANP) algorithm [27].

Biological Modifiers

The original motivation for SplicerAV was to identify statistically significant changes in alternative processing that would also provide ideal targets for further experimental validation and study. To this end, we incorporated additional, user-modifiable parameters, which can preferentially rank events expected to be more amenable to experimental investigation. There are three biological modifiers applied to the MLR to generate the final splice score: a multiple probeset correction to adjust for total possible paired groupings of probesets, an expression cutoff modifier to specify the minimum change required between isoforms, and a centering modifier to preferentially rank genes whose probeset expression levels change in opposite directions. All modifiers are normalized by the average number of paired control and treatment observations for all probesets within a gene (Avg_Obs), so that large samples with higher statistical power will be as influenced by the modifiers as smaller samples, providing parameters that can be applied with consistent effects across varying sample sizes (see equation 2 and methods).

$$\begin{aligned} Splice\ Score &= MLR + Avg\_Obs \\ &* (Multiple\ probeset + Cutoff + Centering\ Modifiers) \end{aligned} \quad (2)$$

These modifiers do not affect the p-value generated by SplicerAV, but allow the program to preferentially rank predicted changes in alternative processing that generate less complicated hypotheses, are larger in magnitude, reflect changes in expression which are qualitatively different, and are less likely to reflect probesets targeting non-transcribed regions or probesets that do not linearly reflect changes in transcript abundance. Genes that exhibit statistically significant discordant probeset behavior and are given a positive splice score represent ideal candidates for experimental investigation of isoform specific regulation.

SplicerAV generates several additional outputs with each file. These include a file containing assessment of statistically significant expression changes for all probesets, a log file containing all user set parameters and comparisons made, as well as a FASTA file for each gene. These fasta files contain the target sequences of all probesets targeting that gene, allowing quick and easy mapping to known and predicted mRNA sequences

using the UCSC genome browser http://genome.ucsc.edu[28]. All genomic analyses in this study were performed using the March 2006 release of the human genome (hg18).

SplicerAV Index Generation

To perform analyses of isoform changes within individual samples we derived an index of relative isoform abundance predicted by SplicerAV. High-throughput analyses of alternative processing have previously defined "splice index" as a quantitative measure to compare isoform abundances between individual samples. The splice index of a probeset equals its expression relative to other probesets targeting the same gene [29]. Using SplicerAV we defined a modified version of the splice index, referred to as the SplicerAV index. SplicerAV assumes a Gaussian mixture model, whereby all probesets are classified as belonging to one of two groups based on similarity of expression changes. The group of probesets exhibiting the largest increases in expression are referred to as the "A" (up) group and the group of probesets exhibiting the largest decreases in expression are referred to as the "B" (down) group (see examples of SplicerAV output in additional files 1, 2, 3, 4, 5, and 6). The SplicerAV index of a probeset equals its expression relative to the average expression of probesets in the opposite group. For example, the SplicerAV index of a probeset in the "A" group would be calculated by subtracting the average expression of the "B" group from that probeset's log2 expression value. In our analysis, SplicerAV indexes of probesets in the "A" group were defined as increased in aggressive cancers, while indexes of probesets in the "B" group were defined as decreased in aggressive cancers. Pre-specified hypotheses generated in training datasets made unidirectional significance tests appropriate in independent validation datasets.

SplicerAV Implementation

SplicerAV was implemented in Perl, with a typical run time of 3-5 minutes on a standard personal computer and has not been tested using other operating systems. The program will only assess changes in alternative mRNA processing for genes interrogated by multiple probesets, which varies widely by microarray platform. To explore the potential for SplicerAV to identify novel changes in mRNA isform abundance in breast cancer, we applied SplicerAV to several publicly available, archival Affymetrix HG-U133 plus 2.0 datasets.

SplicerAV predicts oncogene induced changes in alternative processing of splicing factors

Studies of *SRC* [30], *HRAS* [31,32], and *E2F* family binding sites [33] have demonstrated isolated roles of these oncogenes in affecting alternative mRNA processing. Nonetheless, prior to this study no large-scale examination of changes in alternative mRNA processing had been undertaken for any of these oncogenes. We examined an oncogene over-expression microarray dataset published by Nevins and colleagues [23] (GEO accession GSE3151) to demonstrate SplicerAV's ability to detect oncogene driven changes in alternative processing. In this experiment, activated HRAS, SRC, E2F3, activated β-catenin (CTNNB1), MYC, or green fluorescent protein (GFP) was over-expressed in human primary mammary epithelial cells. The Affymetrix U133 plus 2.0 microarray platform was used to assay gene expression in seven to ten replicates of each condition. Probeset level intensities were estimated using the Robust Multichip Averaging (RMA) procedure [34].

SplicerAV compared changes in probeset expression between GFP and over-expression of the HRAS, SRC, E2F3, CTNNB1, or MYC oncogenes (additional files 1, 2, 3, 4, and 5). Roughly 7,000 genes were expressed above background in either GFP or oncogene over-expression, depending on the oncogene ("Total" column; Table 2). More than 2,000 of these genes were interrogated by multiple probesets, and could therefore be examined by SplicerAV for evidence of changes in alternative mRNA processing ("Multi-probeset Genes" column). More than a hundred isoform specific changes were predicted for each oncogene (Example SplicerAV output shown in Figure 2A; "Alt. Processed Genes" column Table 2). HRAS over-expression caused 645 significant isoform changes, suggesting HRAS-induced changes in alternative processing in nearly a tenth of all expressed genes. The median relative fold change between isoforms was 1.39 ($\log_2$ fold change of .48), with 61 (9%) of these genes predicted to undergo a greater than two fold change in relative isoform abundance (Figure 2B).

Gene isoform changes receiving both a significant p-value and a positive splice score indicate ideal candidates for further experimental study ("Genes with Splice Score > 0" column; Table 2). HRAS and SRC over-expression resulted in 212 and 119 such events, while MYC over-expression resulted in only 12 (Table 2). One gene, Programmed Cell Death Protein 5 (*PDCD5*), underwent the same change in alternative processing upon over-expression of each of the five oncogenes (see additional files 1, 2, 3, 4, and 5). PDCD5 switched from an alternative isoform (mRNA AK293486) to the major isoform (mRNA BC015519), which codes 37 isoform specific c-terminal amino acids required for PDCD5 nuclear entry & activation of apoptosis [35]. Gene ontology (GO) analysis of isoform specific changes revealed a common selection for genes involved in mRNA splicing (see methods). Over-expression of all oncogenes other

Table 2 SplicerAV predicts oncogene-induced changes in isoform specific mRNA levels

	Unique Expressed Genes		SplicerAV Predictions (P < .01)		
GFP vs.	Total	Multi-probeset Genes	Alt. Processed Genes	Genes with Splice Score > 0	Significant Gene Ontologies
HRAS	7227	2185	645	212	**mRNA splicing (12)** Complement med immunity (3) G-protein mediated signaling (10)
SRC	7007	2015	291	119	Transcription Elongation (2) **mRNA splicing (7)**
CTNNB1	7023	2019	159	54	**mRNA processing factors (4)**
E2F3	7313	2139	187	45	Cell surface receptor signal (10) G-protein mediated signaling (6) Mesoderm development (6) Cell structure and motility (11) **pre-mRNA splicing (5)** Granulocyte-mediate immunity (2)
MYC	7081	2040	115	12	—

The total number of unique genes expressed above background, targeted by multiple probesets, predicted to undergo changes in alternative processing, or those predicted to undergo ideal changes are shown in their respective columns for each oncogene (see text). Significantly enriched ($p \leq .05$) biological processes or molecular functions of these genes are shown in order of decreasing significance. The number of genes in each GO category is shown in parentheses.

A

Rank	Gene Symbol	Probeset Name	Log2 Fold Change	P-val	Group	Splice Score	SplicerAV P-val	ANOVA P-val
1	EGFR	210984_x_at	0.662	1.71E-08	A_	923.97	0	0
		201983_s_at	0.541	5.95E-09	A_			
		211607_x_at	0.531	1.63E-07	A_			
		224999_at	0.51	7.90E-08	A_			
		233044_at	0.476	7.00E-03	A_			
		232925_at	0.296	5.60E-02	A_			
		232120_at	-0.039	7.39E-01	_B			
		232541_at	-0.048	6.86E-01	_B			
		201984_s_at	-0.308	1.92E-05	_B			
		1565484_x_at	-0.881	1.79E-04	_B			
		1565483_at	-0.961	2.61E-04	_B			
2	JMJD1C	224933_s_at	1.017	5.23E-11	A_	550.32	0	0
		228793_at	0.823	9.24E-09	A_			
		221763_at	0.731	1.14E-04	A_			
		241661_at	-1.164	1.97E-12	_B			
3	MMP28	219909_at	0.454	3.95E-06	A_	437.61	0	0
		222937_s_at	0.432	4.82E-05	A_			
		224207_x_at	0.19	3.57E-04	A_			
		239273_s_at	0.152	7.37E-02	A_			
		239272_at	-0.691	1.07E-05	_B			

B

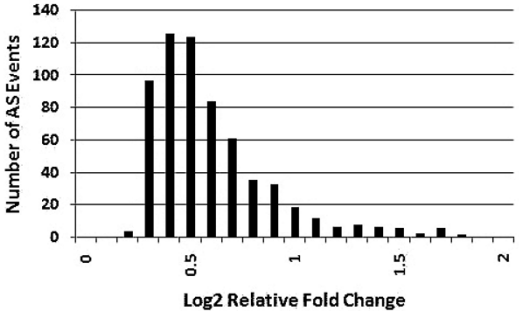

Figure 2 HRAS over-expression results in substantial relative isoform changes. (A) Example SplicerAV output comparing HRAS to GFP over-expression. Genes are ranked in order of descending Splice Score (top three genes shown), with EGFR receiving the top score in HRAS over-expression. Log_2 fold change in expression and corresponding p-values from two tailed homoskedastic t-test of differential expression are shown for individual probesets targeting each gene. Probesets are placed into A and B groupings by SplicerAV (see text). Splice score, SplicerAV p-value, and two way ANOVA p-values are shown for each gene. (B) Distribution of the 645 isoform changes (AS Events) predicted by SplicerAV ($p < .01$) upon HRAS over-expression in human primary mammary epithelial cells. For each gene, SplicerAV separates probesets into two similarly behaving groups based on similar fold changes in expression. The average change in expression between probesets in these two groups (AvgChange, see Equation 8 in methods) reflects the relative fold change in isoform abundance predicted by SplicerAV. Absolute relative fold change in isoform abundance is shown in log base 2.

than MYC each resulted in significant ($p \leq .05$) enrichment of isoform specific changes in mRNA splicing, pre-mRNA splicing, or mRNA processing factors (Table 2). HRAS and SRC over-expression resulted in predicted isoform changes in 12 ($p = .009$) and seven ($p = .05$) factors involved in mRNA splicing, respectively. Both HRAS and E2F3 isoform specific changes were enriched for G-protein mediated signaling ($p = .04$; $p = .0009$) and roles in immune function ($p = .02$; $p = .01$). Sixty-seven genes were predicted to undergo isoform changes in common between two or more oncogenes. Messenger RNA processing factors (5 genes, $p = .008$; *WDR33, HNRPC, SF3A1, SNRPA1, TRA2A*) and mRNA splicing factors (8 genes, $p = .0003$; *HNRPC, HNRPD, TARDBP, HNRPH1, SF3A1, HNRPA2B1, SNRPA1, TRA2A*) were the most significant molecular function and biological process represented by these genes.

HRAS over-expression results in isoform specific EGFR mRNA regulation

Epidermal growth factor receptor (*EGFR*) was the top ranked gene prediction in HRAS over-expression ($p < 10^{-5}$; additional file 1: Tab delimited SplicerAV output of HRAS vs. GFP over-expression). *EGFR* expression was interrogated by seven probesets, providing an ideal opportunity to examine the behavior of multiple probesets targeting different regions of the same gene. Depending on the *EGFR* region being interrogated, probesets reported either a significant increase or decrease in expression upon HRAS over-expression (Figure 3). Four main mRNA isoforms of *EGFR* are annotated in the NCBI database, labeled A, B, C, and D. Isoform A encodes the full length membrane bound tyrosine kinase receptor [36,37]. Variants of isoform A have been observed with either long (A_{Long}) or short (A_{Short}) 3'UTRs (UCSC mRNA accession X00588[36] and AK225422 [38]). Isoforms B and D encode truncated intracellular domains (RefSeq NM_201282; RefSeq NM_201284) and isoform C (RefSeq NM_201283) encodes an *EGFR* variant that lacks a trans-membrane domain and is expected to be soluble [39].

Probesets 1 and 2, which target a region common to all four isoforms, reported highly concordant ($R^2 = .95$) expression levels across all 55 samples in the dataset (Figure 3C). Probesets targeting different transcript regions (1 and 3) reported poor or even inversely correlated expression levels, ($R^2 = .36$, Figure 3D). Due to this "outlier" behavior these probesets would be discarded during conventional microarray expression analysis [40], however, SplicerAV data suggest that this behavior reflects isoform-specific regulation of *EGFR* expression

EGFR isoform A (A_{Short}) appeared to be the primary transcript upregulated by HRAS over-expression, as evidenced by highly correlated expression of the probesets targeting the common and A_{Short} isoforms (probesets 1 and 6; $R^2 = .87$). HRAS over-expression caused a robust decrease in the probeset targeting the long 3'UTR of *EGFR* (probeset 7; A_{Long}) that was not correlated with expression of the common transcript region (Figure 3F, $R^2 = .01$). In contrast, common and A_{Long} expression levels were well correlated in non-HRAS samples ($R^2 = .70$). These data suggest a HRAS-specific shortening of the isoform A 3'UTR.

We hypothesize that these HRAS-induced isoform changes promoted *EGFR* activation via several mechanisms. HRAS increased overall isoform A transcript levels, as evidenced by significant increases in probesets interrogating common regions of the gene (probesets 1 & 2). At the same time, HRAS over-expression resulted in selection of a shorter 3' UTR, which removes known miRNA binding sites present in the A_{Long} UTR and likely increased translation of *EGFR* mRNAs [41]. Widespread 3'UTR shortening to escape miRNA regulation has been observed previously in proliferating cells [42]. *EGFR* isoforms B & D code for a truncated intracellular domain, which if translated could dimerize with and inhibit activation of both *EGFR* and *HER2* [37]. The observed down-regulation of these isoforms is predicted to promote *EGFR1* and *HER2* activation [37]. It should be noted, however, that the corresponding truncated receptors have not been observed. Soluble isoforms composed of the extracellular domain occur naturally and suppress ligand-dependent *EGFR* signaling and oncogenic transformation in a dominant negative manner [43]. Our data indirectly address expression levels of the soluble isoforms, which appear to be unchanged.

Our data suggest that HRAS acts through several isoform-specific mechanisms to promote *EGFR* family signaling. *EGFR* signaling plays known roles in cell survival, proliferation, adhesion, migration, and differentiation [44]. Both *EGFR* and *HER2* are currently therapeutic targets in breast cancer [45]. Our analysis here suggests that modified regulation of alternative mRNA processing could be used as a novel means of *EGFR* inhibition, similar to that shown recently for *HER2* using splice site switching oligonucleotides [46].

SplicerAV predicted isoform changes exhibit low overlap with gene expression changes

Using the same gene expression dataset, SplicerAV was able to predict a number of previously unappreciated changes in isoform specific mRNA regulation. Genes predicted to undergo isoform changes exhibited small overlap with genes predicted to undergo expression changes by conventional analysis, consistent with previous findings in the field [1,47,48]. HRAS and SRC over-expression resulted in the largest changes in both gene expression and isoform changes. Of the212 genes

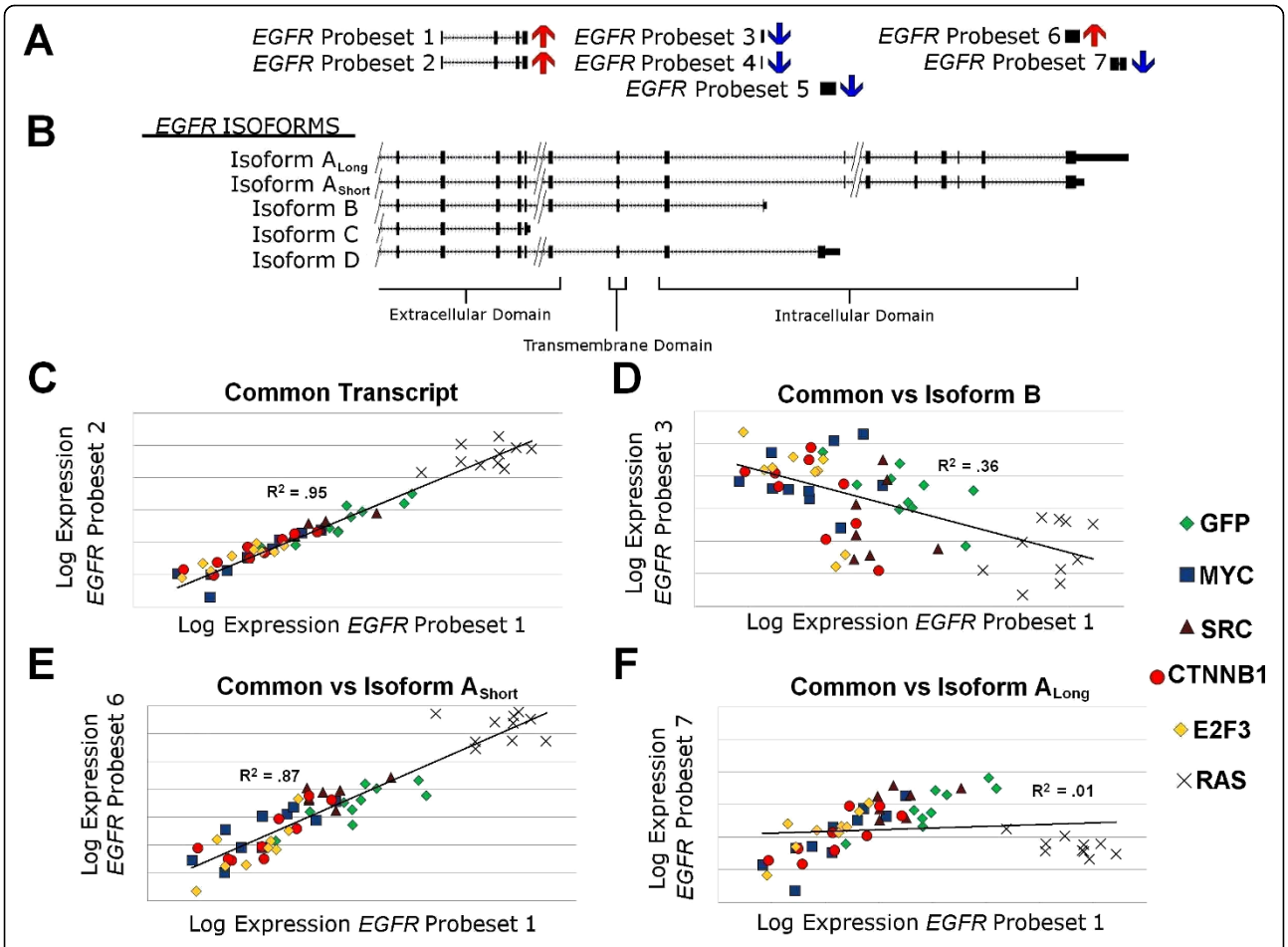

Figure 3 HRAS over-expression causes isoform specific regulation of Epidermal Growth Factor Receptor (EGFR) in human mammary epithelial cells. (A) Probesets on the Affymetrix U133 2.0 plus array interrogate EGFR expression at seven different genomic locations. Up and down arrows indicate each probeset's expression changes in HRAS over-expression compared to GFP controls. Probeset 5 experienced a significant decrease in expression with HRAS over-expression, but was not expressed above background. B) UCSC genomic alignment of probesets and EGFR isoforms. Four previously observed EGFR isoforms (A, B, C and D) are shown with exons represented as black boxes and introns as hashed lines. Extracellular, transmembrane, and intracellular domain regions are shown below the alignment. C-F) Scatter plots of logged expression levels of all 55 samples (GFP, MYC, SRC, CTNNB1, E2F3, and HRAS) for selected pairs of probesets C) Probesets 1 and 2 target a transcript region common to all major isoforms and exhibit highly correlated expression (R2 = .95). D) Probesets 1 and 3 target the common region vs. isoform B specific region and demonstrate a weak inverse relationship (R2 = .36). E) Probesets 1 and 6 interrogate the common vs. AShort isoform region, demonstrating a high degree of correlation across all samples (R2 = .87). F) In contrast, probesets 1 and 7 interrogate common and ALong isoform region and are not correlated (R2 = .01) due to the HRAS induced 3'UTR shortening of EGFR A transcripts.

predicted to undergo ideal isoform changes (significant p-value and positive splice score) in HRAS over-expression, only 8 genes (3.8%) were also among the top 212 most significant changes by conventional expression analysis (data not shown). Of the top 119 predicted isoform changes in SRC over-expression, none were in the top 119 most significant expression changes. This low degree of overlap suggests that the results obtained via SplicerAV are largely orthogonal to that of conventional gene expression analyses. This low degree of overlap provides the potential for combining traditional gene expression signatures with SplicerAV isoform-based signatures to improve signature performance.

SplicerAV predicts isoform changes in high vs. low grade breast tumors

Our analysis of oncogene regulated isoform expression demonstrated the ability to generate novel insights into cancer biology. We next determined if similar insights could be obtained from the analysis of alternative processing in clinical tumor samples. Breast cancer has been extensively studied using high-throughput analyses of gene expression at the transcriptome level (Reviewed in [49]). In contrast, high-throughput analysis of alternative mRNA processing in breast cancer has been addressed in only a handful of studies [12,47]. We explored the ability of SplicerAV to detect changes in

alternative processing between low and high grade breast tumors in archival expression data.

Sotiriou and colleagues profiled 87 Tamoxifen treated, estrogen receptor (ER) positive tumors obtained from Guys Hospital, London (GUYT) using the Affymetrix HG-U133 PLUS2 Genechip™[50] (GEO accession GSE6532, RMA normalized). Using this dataset, we examined changes in probeset expression between low grade (I, n = 17) and high grade (III, n = 16) breast tumors. Analysis was limited to probesets present on either the U133A or U133B arrays in order to validate changes in two independent data sets discussed in the next section. 11,248 unique genes were expressed above background in either the low or high grade tumor samples. Among the 4,031 genes interrogated by multiple probesets, SplicerAV predicted that 974 genes underwent significant isoform changes between aggressive and non-aggressive breast tumors ($p < .01$; see additional file 6: Tab delimited SplicerAV output of Grade I vs. Grade III human breast tumors). Removing genes with negative splice scores yielded a refined list of 241 genes. GO analyses of these 241 genes revealed significant ($p < .05$) enrichment for several molecular functions including guanyl-nucleotide exchange factors (*RAB3IP, RAPGEF2, GAPVD1, CD47, TRIO, ARHGEF7, AKAP13*; p = .006), metalloprotease inhibitors (*TIMP2, TIMP3*; p = .007), ubiquitin-protein ligases (*RNF130, TTC3 UBE3B, PML, TRIM26, RBCK1, MIB1, ZNF294, ZUBR1, TRIAD3*; p = .007), and mRNA processing factors (*SYNCRIP, WDR33, SFRS8, SFRS15, TAF15, SF1, SF3B1, SFPQ, PRP6*; p = .01; Table 3).

SplicerAV predicted isoform changes are associated with breast cancer survival

SplicerAV probeset groupings of genes identified in the GUYT training set were used to create individual sample level indexes of relative isoform abundance. We tested an association of these SplicerAV indexes in two independent validation datasets to examine whether specific isoform changes observed in high grade tumors were also associated with poor patient prognosis (see methods). Previous datasets generated by Miller [51] (GSE3494) and Pawitan [52] (GSE1456) have independently profiled breast tumor gene expression using the Affymetrix U133 A and B microarrays (probeset intensities were estimated using MAS5 [53]). These studies include patient outcome, providing the opportunity to test for an association of isoform changes with survival in ER positive tumors.

We generated 687 SplicerAV Indexes from the 241 genes identified in the GUYT training set and calculated their value for each tumor sample in the validation sets. For each SplicerAV Index, tumors were sorted into the top and bottom 50^{th} percentile of tumors. High and low SplicerAV Index groups were then tested for a difference in survival. The GUYT training set had previously determined whether a SplicerAV index was predicted to be increased or decreased in aggressive cancer (defined as Grade III vs Grade I). This pre-specified association with aggressive cancer was used to conduct one-sided logrank tests ($p < .05$) for an association with breast cancer survival for each SplicerAV index in the validation datasets. Failure in the Miller dataset was defined as death from any cause and failure in the Pawitan dataset was defined as death from breast cancer (inherent to the clinical data available). Of the 241 genes tested, 15 genes possessed indexes that were significantly associated with survival in both datasets (Table 4). Guanyl-nucleotide exchange factors (GEFs) and mRNA processing factors were both enriched among the original 241 genes tested. Interestingly, these GO categories were both represented among the 15 validated genes including *ARHGEF7*, a guanyl-nucleotide exchange factor, and *SFPQ*, an mRNA processing factor.

Few studies have performed high-throughput examination of alternative processing in clinical tumor samples [12,13] and to our knowledge no prior studies have examined changes in alternative mRNA processing

Table 3 GO analysis of 241 genes predicted to undergo isoform changes between grade I and grade III breast tumors (GUYT)

Molecular Function	# Genes	P-Value	Gene Symbols
Guanyl-nucleotide exchange factor	7	6.22E-03	*RAB3IP, RAPGEF2, GAPVD1, CD47, TRIO, ARHGEF7, AKAP13*
Metalloprotease inhibitor	2	6.52E-03	*TIMP2, TIMP3*
Ubiquitin-protein ligase	10	7.40E-03	*RNF130, TTC3 UBE3B, PML, TRIM26, RBCK1, MIB1, ZNF294, ZUBR1, TRIAD3*
mRNA processing factor	9	1.27E-02	*SYNCRIP, WDR33, SFRS8, SFRS15, TAF15, SF1, SF3B1, SFPQ, PRP6*
Cytoskeletal protein	4	3.42E-02	*DNAL1, NF2, KIF5C, DYNC1H1*
Anion channel	2	3.63E-02	*PML, CLCN3*
G-protein modulator	12	4.64E-02	*RAB3IP, RAPGEF2, GAPVD1, CD47,*
mRNA splicing factor	6	4.94E-02	*TAF15, SFRS8, SF1, SF3B1, SFPQ, PRP6*
Tyrosine protein kinase receptor	4	4.97E-02	*TEK, TPR, IGF1R, PDGFRA*

Table 4 Isoform changes in gene expression significantly associated with patient outcomes in both validation datasets

SplicerAV Predictions			Association with Survival	
Gene Symbol†	Isoform Probeset	Hypothesis	Miller	Pawitan
ARHGEF7	202548_s_at	DOWN	*0.009	*0.008
DPP7	241973_x_at	DOWN	*0.001	*0.007
EIF4E2	209393_s_at	UP	**0.002	*0.003
MAPKAP1	222426_at	DOWN	*0.019	*0.003
SLC28A10	230448_at	UP	*0.007	0.032
PDXK	202671_s_at	UP	**0.001	0.025
POLI	238992_at	UP	0.037	0.052
SFPQ	201585_s_at	UP	0.062	0.041
SIVA1	203489_at	UP	*0.005	0.075
SSU72	223051_at	UP	*0.018	*0.007
TFDP2	203588_s_at	UP	0.054	*0.008
TIMP2	231579_s_at	DOWN	**0.001	0.056
TncRNA	234989_at	UP	**0.001	0.034
WDFY3	212606_at	UP	0.049	*0.010
WDR26	224897_at	UP	**0.001	0.049

†For genes possessing multiple significant SplicerAV Indices, only one isoform is shown.

*Significant association with survival (p < .01), one sided log rank test

** Significant association with survival (p < .001), one sided log rank test

directly associated with cancer patient survival. This study examined isoform specific mRNA levels in over 400 human clinical samples, providing support for the use of changes in alternative processing as potential prognostic markers in cancer.

ARHGEF7 &EIF4E2 isoform changes are associated with breast cancer survival

A SplicerAV index for Rho guanine nucleotide exchange factor 7 (*ARHGEF7*) was decreased in high vs. low grade tumors within the GUYT dataset, and was significantly associated with survival in both the Miller (p = .008) and Pawitan (p = .009) datasets. *ARHGEF7* expression was assayed by three annotated probesets, providing an opportunity to compare associations of survival with either SplicerAV index or individual probeset expression. The SplicerAV index for *ARHGEF7* compared the ratio of a decreasing ("Down") probeset located in the 3'UTR of *ARHGEF7* to that of two increasing ("Up1" and "Up2") probesets located in shorter transcripts (Figure 4A). We compared the *ARHGEF7* SplicerAV index and each individual probeset for an association with breast cancer survival and noted that the SplicerAV index outperformed individual probeset in both datasets (Figure 4B).

A SplicerAV index for Eukaryotic translation initiation factor 4E family member 2 (*EIF4E2*) was increased in

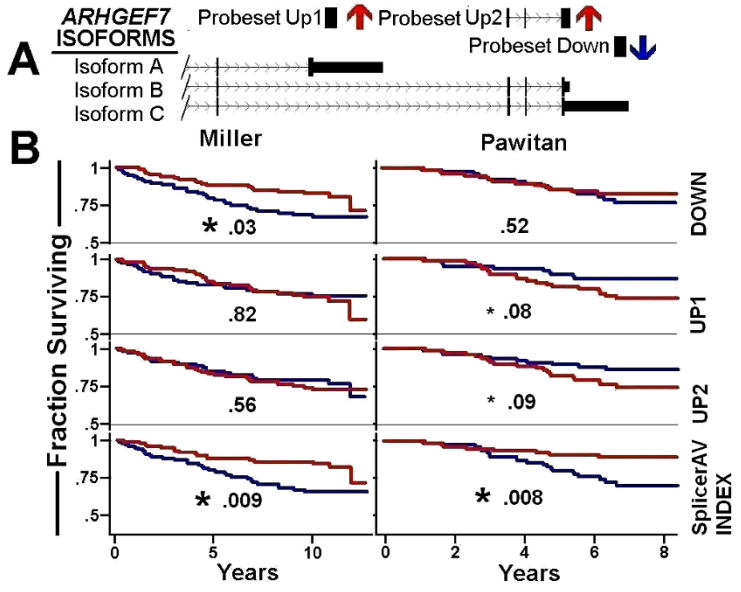

Figure 4 SplicerAV Index of ARHGEF7 is associated with breast cancer survival. Panel A. Schematic representation of ARHGEF7 isoforms A, B and C, with regions interrogated by probesets that increase shown as Probesets Up 1 and 2 (red arrows), and the region which decreases denoted as Probeset Down (blue arrow). Panel B. The fraction of patients surviving in each cohort (vertical axis) is shown over time in years (horizontal axis) as a function of individual probeset expression or SplicerAV index. Survival of patients in the top (red line) and bottom (blue line) 50th percentile are plotted by individual probeset expression (Down, UP1, and UP2) and the SplicerAV index within the Miller (left) and Pawitan (right) cohorts. Results of two-tailed logrank tests of survival are shown, with asterisks indicating significance at the .05 (large asterisk) and .10 (small asterisk) levels.

high vs. low grade tumors within the GUYT dataset, and was significantly associated with survival in both the Miller (p = .002) and Pawitan (p = .003) datasets. The SplicerAV index for *EIF4E2* compared the ratio of an increasing "Up" probeset targeting a coding region to that of a decreasing "Down" probeset located in the 3'UTR of the longest transcript (Figure 5A). For *EIF4E2*, survival could be predicted by an increase in the "Up" probeset alone (Miller, p = .003; Pawitan, p = .0007; Figure 5B). Low levels of the "Down" probeset were only significantly associated with poor survival in the Pawitan cohort (p = .04).

Whether or not individual probesets could demonstrate a consistent association with survival differed by gene. Although individual probeset behavior may represent an alternative processing event, only through comparison with other probesets for that gene can SplicerAV uncover these relevant and predictive isoforms that would go unnoticed in conventional analyses.

Combining isoform changes from multiple genes improves prediction of breast cancer survival

We chose a subset of the 15 validated isoform changes to examine the potential for generating an isoform signature that combined information from multiple isoform changes to improve prognostic accuracy. We initially chose the six genes, *EIF4E2, ARHGEF7, SLC28A10, PDXK, TncRNA*, and *MAPKAP1*, that produced the clearest separation between good and poor survival in individual prognostic analyses (data not shown). Stratifying patients by SplicerAV index for each gene demonstrated the expected association with survival (Figure 6A-F). The number of poor prognostic events was tallied for each patient. Survival was then plotted for individuals with low (0-1 events, blue), intermediate (2-4 events, black), or high (5-6 events, red) numbers of poor prognostic events (Figure 6G). This stratification of patients by total poor prognostic events demonstrated highly significant associations with survival in both the Miller (p = 6e-7) and Pawitan (p = 4e-7) cohorts. The combined isoform signature demonstrated prognostic value beyond that of any individual isoform or probeset change.

Similar to our *in vitro* analyses of oncogene overexpression, we observed low overlap between gene expression and SplicerAV changes. Of the 241 isoform changes predicted by SplicerAV in the GUYT training set that were later tested for an association with poor prognosis, only one gene (0.4%), *BTD*, was also among the top 241 differentially expressed genes. The orthogonality of candidate gene lists identified by SplicerAV and conventional methods suggests that these two

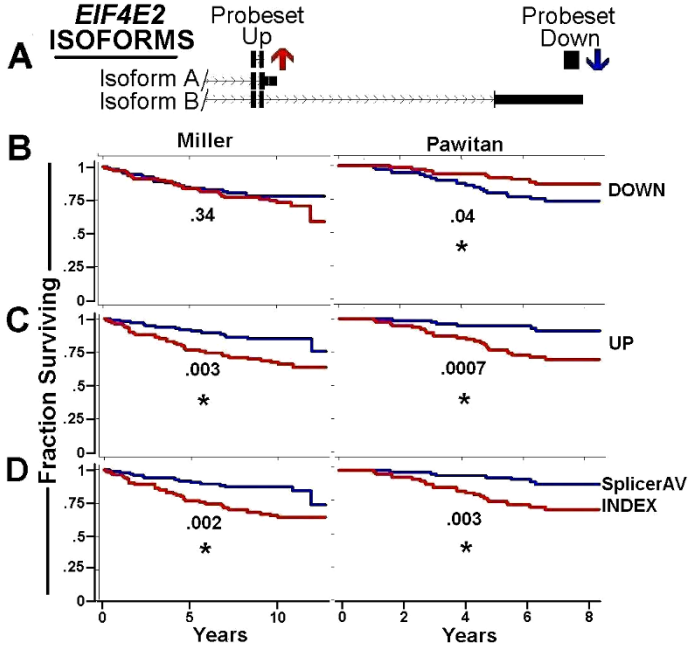

Figure 5 EIF4E2 probesets are associated with breast cancer survival. Panel A. Schematic representation of EIF4E2 isoforms A and B, with region interrogated by probesets shown as Up (red arrow), and Down (blue arrow). For panels B, C, and D, the fraction of patients surviving in each cohort (vertical axis) is shown over time in years (horizontal axis) as a function of individual probeset expression or SplicerAV index. Survival of patients in the top (red line) and bottom (blue line) 50th percentile are plotted by individual probeset expression (B, C) and the SplicerAV index (D) within the Miller (left) and Pawitan (right) cohorts. Results of two-tailed logrank tests of survival are shown, with asterisks indicating significance at the .05 level.

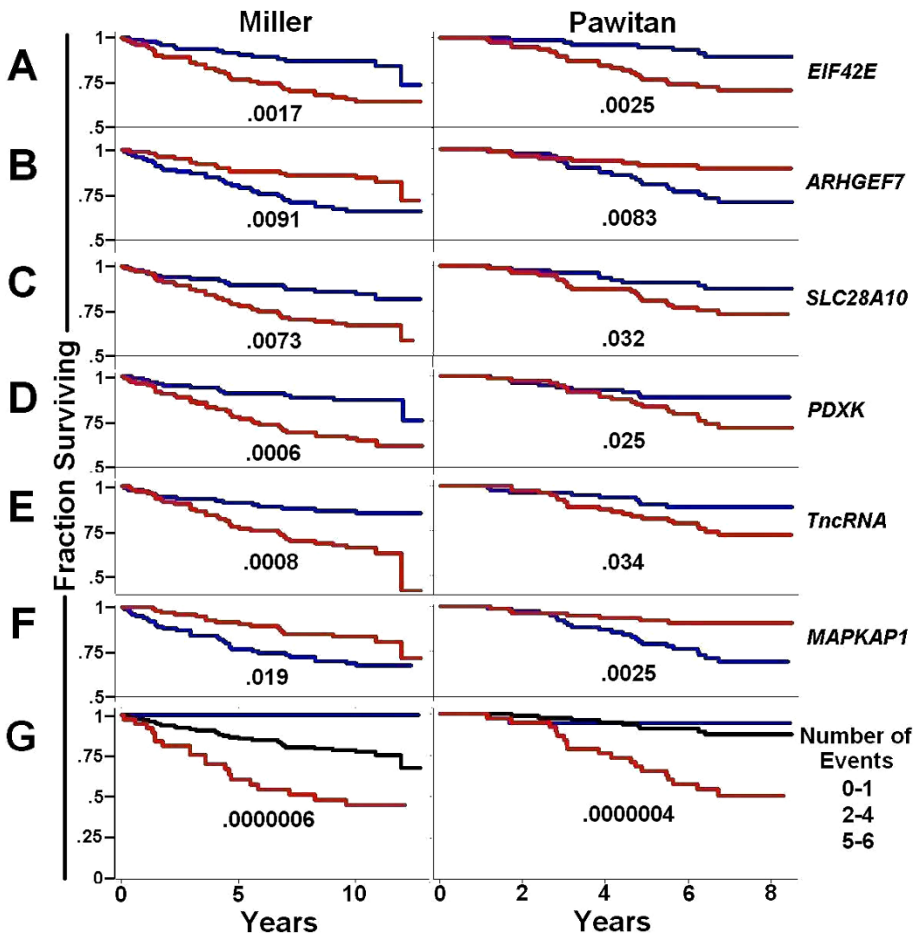

Figure 6 A six isoform signature provides improved prediction of breast cancer survival compared to individual isoforms. The fraction of patients surviving in each cohort (vertical axis) is shown over time in years (horizontal axis) as a function of individual probeset expression or SplicerAV index. Survival of patients in the top (red line) and bottom (blue line) 50th percentile are plotted by the SplicerAV index for six genes; EIF4E2 (A), ARHGEF7 (B), SLC28A10 (C), PDXK (D), TncRNA (E), MAPKAP1 (F) for the Miller (left) and Pawitan (right) cohorts. Patients survival stratified by a low (0-1), intermediate (2-4), and high (5-6) number of poor prognostic events is shown in panel G.

methods detect different biological processes and may provide independent value in generating molecular classifiers. SplicerAV can generate both conventional and isoform specific gene expression analyses, and therefore provides two non-redundant datasets from one experiment.

General Discussion

Traditional analyses of gene expression data have considered the probeset as the basic unit of expression. Under this paradigm, the presence of multiple probesets has been viewed largely as a nuisance. Current approaches dealing with the issue of multiple probesets have used either probeset location or the mean, median, or largest probeset expression change to distill multiple probesets into a single gene level expression value. Each of these approaches would have yielded a different readout of *EGFR* expression changes in HRAS overexpression, making conventional interpretation inadequate for such genes. Software has even been developed whose sole purpose is the removal of discordant probeset expression values for probesets targeting the same gene [40].

We propose that for genes with multiple probesets, isoform specific expression changes may be a more appropriate means of interpreting standard microarray expression data than the current one gene = one probeset paradigm. Previous algorithms [54,55] have examined the possibility of investigating changes in alternative processing using single probeset level data. These methods have relied on custom chips, or would not have detected events predicted by SplicerAV in this paper because such methods do not examine events spanning multiple probesets. SplicerAV provides a systematic means by which to detect and interpret inconsistent probeset behavior within the same gene, a

situation where an oversimplified perspective may be obscuring relevant and important biological changes.

This study marks the first *en masse* analysis of mRNA isoform changes in existing conventional expression microarray data. We have shown here that re-analyzing such data using a different paradigm can uncover novel biological insights and potential prognostic markers.

Conclusion

The combination of material, personnel, and clinical costs of obtaining gene expression microarray data has resulted in a massive archive of these data accumulated over the past two decades. Many previously created datasets, particularly clinical datasets, are unique and cannot be reproduced. Numerous private and public repositories of microarray expression data exist, with the largest public repository, Gene Expression Omnibus, containing over 50,000 data samples from the Affymetrix U133 and U95 series alone. In this paper we demonstrate the utility of SpicerAV, the first program used to analyze this existing data *en masse* for isoform specific changes that can result from alternative mRNA processing.

Methods

SplicerAV algorithm details

SplicerAV takes probeset intensities generated using conventional normalization methods (i.e. MAS5 or RMA output) as input. SplicerAV first summarizes the average $\log_2$fold change in expression and the corresponding t-statistic for each probeset on the array. Probeset changes are assigned an initial weight based on their normalized t-statistic, T_{Norm}. Conceptually, weighting by T_{Norm} counts probesets undergoing significant expression changes one time. This is because T_{Norm} equals one for probesets reporting expression changes significant at the .05 level (two tailed t-test).

$$T_{Norm} = \frac{|\mu_{Treatment} - \mu_{Control}|}{\sqrt{\sigma^2_{Treatment}\ \sigma^2_{Control}}} \div T_{Critical} \quad (3)$$

Probesets targeting the same gene are next grouped together using annotation provided by the array manufacturer. Genes targeted by probesets with a T_{Norm} value greater than one scale their weights so that the maximum T_{Norm} within that gene is reduced to one. This prevents counting any probeset more than once.

$$If\ Max(T_{Norm}) > 1\ then\ Weight = \frac{T_{Norm}}{Max(T_{Norm})} \quad (4)$$

$$Else\ Weight = T_{Norm}$$

At this step, individual probeset weights are raised to a user specified power (*Wt_scale*, default = 2), which allows preferential focus on more significant probeset changes in expression at the cost of removing information from less reliable probesets and reducing the power of significance tests.

This weighting scheme assigns a weight between 0 and 1 to each probset, indicating the number of times a probeset's observations will be counted in the Gaussian mixture model. In the final Gaussian mixture model, each probeset weight is multiplied by the average number of paired observations among treatment and control groups for that probeset ($N_{avg\_obs} = (N_{treat\_obs} + N_{control\_obs})/2$). The resulting model counts each effective pair of observations for a probeset at most once, with less reliable probesets being counted less.

$$\begin{aligned} &Effective\ Weight_{prbset} \\ &= EfWt_{prbset} \\ &= (Weight_{prbset})^{Wt\_Scale} * Avg\_Obs_{prbset} \end{aligned} \quad (5)$$

The Effective Weight for each probeset is used as the final probeset summary weight in the Gaussian mixture model. Average probeset log2fold changes in expression are fitted using two models, which contain one and two Gaussian distributions, respectively. Comparison of the relative fit under these two models yields a maximum likelihood ratio (MLR), which can be assessed for statistical signifance using a standard likelihood ratio (LR) test statistic, asymptotically distributed as $\chi^2(2)$, for each gene.

$$MLR = \prod_{prbset=1}^{Tot\_prbsets} \frac{Likelihood_A{}^{*IfA} * Likelihood_B{}^{*IfB}}{Likelihood_{Single}} \quad (6)$$

Where:

$$Likelihood_i = \left[\frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(X_{prbset} - \mu_i)^2}{2\sigma_i^2} \right) \right]^{EfWt_{prbset}}$$

X_{prbset} = the log2fold expression change of that probeset

μ_A = the weighted average log2fold change in expression for probesets assigned to groupA

μ_B = the weighted average log2fold change in expression for probesets assigned to groupB

μ_{Single} = the weighted average log2fold change in expression for all probesets targeting the gene

σ_A, σ_B, and σ_{single} for groups A, B, and all probesets are determined by expectation maximization, bounded by a minimum value of 10% to prevent over-fitting by the model. The value of 10% was chosen as a

conservative limit based on empirical observations of summarized significant log2fold probeset changes, which consistently exhibited standard deviations (σ) below 10% across analyzed datasets (data not shown).

Biological Modifiers

SplicerAV incorporates biologically motivated modifiers to alter the relative ranking of potential changes in alternative processing to suit the final objectives of the user. These modifiers can be adjusted by the user and do not affect the p-values reported by SplicerAV. The specified form and magnitude of these biologically motivated modifiers were empirically derived through analysis of several datasets.

Multiple Probeset Modifier

The multiprobeset modifier adjusts the splice score by the total possible ways that all the probesets targeting a given gene can be placed into groups of two. This method penalizes genes containing large numbers of probesets capable of generating a large number of alternative processing hypotheses which are difficult to interpret, using a bonferroni multiple hypothesis correction.

$$Multiprobeset\ Modifier = -\ln(2^{tot\_prbsets-1} - 1) \quad (7)$$

Expression Cutoff Modifier

The expression cutoff modifier calculates the $\log_2$ difference in average expression between the two groups of probesets, A and B. Genes whose expression between groups falls below a user specified threshold minimum fold change are penalized using a smoothed function whose steepness is set using a user specified *sharpness* parameter.

$$\begin{aligned} &If\ AvgChange < Cutoff, \\ &\quad Cutoff\ Modifier = Sharpness * \ln(AvChange\ /\ Cutoff) \end{aligned} \quad (8)$$

Centering Modifier

The centering modifier preferentially ranks genes whose probeset expression changes in opposite directions, suggesting a qualitatively different event which cannot be explained by poor annotation of probesets targeting intronic regions, saturated probeset signals, non-hybridizing probesets, or other probeset expression behavior deviating from a linear relationship with transcript abundance. Genes in which both groups of probesets change in the same direction (either both increasing or decreasing) are penalized, while genes containing groups of probesets with mean expression levels moving in opposite directions are given a bonus.

$$\begin{aligned} &If(Mean\Delta_{GrpA} * Mean\Delta Exp_{GrpB}) < 0, \\ &\quad then\ Centering\ Factor = -Centering\ Factor \\ &Centering\ Modifier \\ &= Centering\ Factor * Min(|\ Mean\Delta_{GrpA}\ |, |\ Mean\Delta Exp_{GrpB}\ |) \end{aligned} \quad (9)$$

Gene Ontology Analyses

Gene ontology (GO) analyses compared genes with SplicerAV predicted isoform changes ($p < .01$, splice score > 0) to a reference set of all genes evaluated for isoform changes in each condition using PANTHER [56,57]. Non-overlapping GO categories with more than one gene were reported.

Additional file 1: Tab delimited SplicerAV output of HRAS vs. GFP over-expression.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-108-S1.TXT]

Additional file 2: Tab delimited SplicerAV output of SRC vs. GFP over-expression.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-108-S2.TXT]

Additional file 3: Tab delimited SplicerAV output of E2F3 vs. GFP over-expression.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-108-S3.TXT]

Additional file 4: Tab delimited SplicerAV output of CTNNB1 vs. GFP over-expression.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-108-S4.TXT]

Additional file 5: Tab delimited SplicerAV output of MYC vs. GFP over-expression.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-108-S5.TXT]

Additional file 6: Tab delimited SplicerAV output of Grade I vs. Grade III human breast tumors (GUYT).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-108-S6.TXT]

Abbreviations

GEO: Gene Expression Omnibus; NSMP: Negative Strand Matching Probeset; PANP: Presence-Absence calls with Negative Probesets; MLR: Maximum Likelihood Ratio; LR: Likelihood Ratio; GFP: Green Fluorescent Protein; UTR: Untranslated Region.

Acknowledgements

We thank Joe Nevins, Holly Dressman, Joe Lucas, and Erich Huang for helpful comments on the manuscript and Ashley Chi, Sayan Mukherjee, Uwe Ohler, and Alexander Hartemink for their suggestions and advice during the development of SplicerAV. We acknowledge funding from the NIH grants 5R01-GM63090 (MGB) and 1R01-CA127727 (MGB), the DOD grant GRANT00412169 (TJR) (Predoctoral Traineeship Award), and the SPORE grant 5P50-CA068438-10 (MD).

Author details
[1]Molecular Cancer Biology Program, Duke University Medical Center, Durham, USA. [2]Department of Health Policy and Management, University of North Carolina at Chapel Hill, Chapel Hill, USA. [3]Department of Radiation Oncology, Duke University Medical Center, Durham, USA. [4]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, USA. [5]Department of Medicine, Duke University Medical Center, Durham, USA. [6]Center for RNA Biology, Duke University Medical Center, Durham, USA.

Authors' contributions
JLP and TJR developed the program. TJR designed and implemented the algorithm and conducted the bulk of the analyses. TJR & MAD were responsible for applying SplicerAV to clinical datasets. MD assisted with the selection and analysis plan of clinical datasets and statistical refinement of the algorithm. MGB and JLP assisted with extensive conceptual refinement of the algorithm and analyses, and directed the research. All authors read and approved the final manuscript.

Received: 23 September 2009
Accepted: 25 February 2010 Published: 25 February 2010

References

1. Blencowe BJ: **Alternative splicing: new insights from global analyses.** *Cell* 2006, **126(1)**:37-47.
2. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456(7221)**:470-476.
3. Krawczak M, Reiss J, Cooper DN: **The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences.** *Hum Genet* 1992, **90(1-2)**:41-54.
4. Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R: **Are splicing mutations the most frequent cause of hereditary disease?.** *FEBS Lett* 2005, **579(9)**:1900-1903.
5. Garcia-Blanco MA, Baraniak AP, Lasda EL: **Alternative splicing in disease and therapy.** *Nat Biotechnol* 2004, **22(5)**:535-546.
6. Venables JP: **Unbalanced alternative splicing and its significance in cancer.** *Bioessays* 2006, **28(4)**:378-386.
7. Cooper TA, Wan L, Dreyfuss G: **RNA and disease.** *Cell* 2009, **136(4)**:777-793.
8. Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, Jin L, Motono C, Hata H, Isogai T, Nagai K, *et al*: **Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs.** *Nucleic Acids Res* 2006, **34(14)**:3917-3928.
9. Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Inkel L, Durand M, Couture S, Froehlich U, Lapointe E, *et al*: **Cancer-associated regulation of alternative splicing.** *Nat Struct Mol Biol* 2009, **16(6)**:670-6.
10. Xu Q, Lee C: **Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences.** *Nucleic Acids Res* 2003, **31(19)**:5635-5643.
11. He C, Zhou F, Zuo Z, Cheng H, Zhou R: **A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis.** *PLoS ONE* 2009, **4(3)**:e4732.
12. Andre F, Michiels S, Dessen P, Scott V, Suciu V, Uzan C, Lazar V, Lacroix L, Vassal G, Spielmann M, *et al*: **Exonic expression profiling of breast cancer and benign lesions: a retrospective analysis.** *Lancet Oncol* 2009, **10(4)**:381-390.
13. Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, *et al*: **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 2006, **7**:325.
14. Xi L, Feber A, Gupta V, Wu M, Bergemann AD, Landreneau RJ, Litle VR, Pennathur A, Luketich JD, Godfrey TE: **Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer.** *Nucleic Acids Res* 2008, **36(20)**:6535-6547.
15. Thorsen K, Sorensen KD, Brems-Eskildsen AS, Modin C, Gaustadnes M, Hein AM, Kruhoffer M, Laurberg S, Borre M, Wang K, *et al*: **Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis.** *Mol Cell Proteomics* 2008, **7(7)**:1214-1224.
16. Cheung HC, Baggerly KA, Tsavachidis S, Bachinski LL, Neubauer VL, Nixon TJ, Aldape KD, Cote GJ, Krahe R: **Global analysis of aberrant pre-mRNA splicing in glioblastoma using exon expression arrays.** *BMC Genomics* 2008, **9**:216.
17. Laajala E, Aittokallio T, Lahesmaa R, Elo LL: **Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies.** *Genome Biol* 2009, **10(7)**:R77.
18. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, *et al*: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, , **37 Database:** D885-890.
19. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL: **Multiplexed biochemical assays with biological chips.** *Nature* 1993, **364(6437)**:555-556.
20. Pearson JL, Robinson TJ, Munoz MJ, Kornblihtt AR, Garcia-Blanco MA: **Identification of the cellular targets of the transcription factor TCERG1 reveals a prevalent role in mRNA processing.** *J Biol Chem* 2008, **283(12)**:7949-7961.
21. Stalteri MA, Harrison AP: **Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips.** *BMC Bioinformatics* 2007, **8**:13.
22. D'Mello V, Lee JY, MacDonald CC, Tian B: **Alternative mRNA polyadenylation can potentially affect detection of gene expression by affymetrix genechip arrays.** *Appl Bioinformatics* 2006, **5(4)**:249-253.
23. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, *et al*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439(7074)**:353-357.
24. Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S: **Novel definition files for human GeneChips based on GeneAnnot.** *BMC Bioinformatics* 2007, **8**:446.
25. Lu J, Lee JC, Salit ML, Cam MC: **Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays.** *BMC Bioinformatics* 2007, **8**:108.
26. Yu H, Wang F, Tu K, Xie L, Li YY, Li YX: **Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data.** *BMC Bioinformatics* 2007, **8**:194.
27. Warren P, Taylor D, Martini PGV, Jackson J, Bienkowska J: **PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays.** *Proc 2007 IEEE 7th International Symposium on BioInformatics & BioEngineering, Cambridge, USA* 2007, 108-115.
28. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12(6)**:996-1006.
29. Srinivasan K, Shiue L, Hayes JD, Centers R, Fitzwater S, Loewen R, Edmondson LR, Bryant J, Smith M, Rommelfanger C, *et al*: **Detection and measurement of alternative splicing using splicing-sensitive microarrays.** *Methods* 2005, **37(4)**:345-359.
30. Neel H, Gondran P, Weil D, Dautry F: **Regulation of pre-mRNA processing by src.** *Curr Biol* 1995, **5(4)**:413-422.
31. Chandler LA, Ehretsmann CP, Bourgeois S: **A novel mechanism of Ha-ras oncogene action: regulation of fibronectin mRNA levels by a nuclear posttranscriptional event.** *Mol Cell Biol* 1994, **14(5)**:3085-3093.
32. Chandler LA, Bourgeois S: **Posttranscriptional down-regulation of fibronectin in N-ras-transformed cells.** *Cell Growth Differ* 1991, **2(8)**:379-384.
33. Darville MI, Rousseau GG: **E2F-dependent mitogenic stimulation of the splicing of transcripts from an S phase-regulated gene.** *Nucleic Acids Res* 1997, **25(14)**:2759-2765.
34. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4(2)**:249-264.
35. Yao H, Xu L, Feng Y, Liu D, Chen Y, Wang J: **Structure-function correlation of human programmed cell death 5 protein.** *Arch Biochem Biophys* 2009, **486(2)**:141-149.
36. Ullrich A, Coussens L, Hayflick JS, Dull TJ, Gray A, Tam AW, Lee J, Yarden Y, Libermann TA, Schlessinger J, *et al*: **Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells.** *Nature* 1984, **309(5967)**:418-425.
37. Kashles O, Yarden Y, Fischer R, Ullrich A, Schlessinger J: **A dominant negative mutation suppresses the function of normal epidermal growth factor receptors by heterodimerization.** *Mol Cell Biol* 1991, **11(3)**:1454-1463.

38. Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S: **Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library.** *Gene* 1997, **200(1-2)**:149-156.
39. Reiter JL, Threadgill DW, Eley GD, Strunk KE, Danielsen AJ, Sinclair CS, Pearsall RS, Green PJ, Yee D, Lampland AL, *et al*: **Comparative genomic sequence analysis and isolation of human and mouse alternative EGFR transcripts encoding truncated receptor isoforms.** *Genomics* 2001, **71(1)**:1-20.
40. Jaksik R, Polanska J, Herok R, Rzeszowska-Wolny J: **Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in Affymetrix microarrays.** *Acta Biochim Pol* 2009, **56(2)**:271-7.
41. Weiss GJ, Bemis LT, Nakajima E, Sugita M, Birks DK, Robinson WA, Varella-Garcia M, Bunn PA Jr, Haney J, Helfrich BA, *et al*: **EGFR regulation by microRNA in lung cancer: correlation with clinical response and survival to gefitinib and EGFR expression in cell lines.** *Ann Oncol* 2008, **19(6)**:1053-1059.
42. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB: **Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites.** *Science* 2008, **320(5883)**:1643-1647.
43. Basu A, Raghunath M, Bishayee S, Das M: **Inhibition of tyrosine kinase activity of the epidermal growth factor (EGF) receptor by a truncated receptor form that binds to EGF: role for interreceptor interaction in kinase regulation.** *Mol Cell Biol* 1989, **9(2)**:671-677.
44. Adamson ED, Wiley LM: **The EGFR gene family in embryonic cell activities.** *Curr Top Dev Biol* 1997, **35**:71-120.
45. Browne BC, O'Brien N, Duffy MJ, Crown J, O'Donovan N: **HER-2 signaling and inhibition in breast cancer.** *Curr Cancer Drug Targets* 2009, **9(3)**:419-438.
46. Wan J, Sazani P, Kole R: **Modification of HER2 pre-mRNA alternative splicing and its effects on breast cancer cells.** *Int J Cancer* 2009, **124(4)**:772-777.
47. Li C, Kato M, Shiue L, Shively JE, Ares M Jr, Lin RJ: **Cell type and culture condition-dependent alternative splicing in human breast cancer cells revealed by splicing-sensitive microarrays.** *Cancer Res* 2006, **66(4)**:1990-1999.
48. Zhang C, Li HR, Fan JB, Wang-Rodriguez J, Downs T, Fu XD, Zhang MQ: **Profiling alternatively spliced mRNA isoforms for prostate cancer classification.** *BMC Bioinformatics* 2006, **7**:202.
49. Sotiriou C, Pusztai L: **Gene-expression signatures in breast cancer.** *N Engl J Med* 2009, **360(8)**:790-800.
50. Loi S, Haibe-Kains B, Desmedt C, Wirapati P, Lallemand F, Tutt AM, Gillet C, Ellis P, Ryder K, Reid JF, *et al*: **Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen.** *BMC Genomics* 2008, **9**:239.
51. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, *et al*: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.** *Proc Natl Acad Sci USA* 2005, **102(38)**:13550-13555.
52. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, *et al*: **Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.** *Breast Cancer Res* 2005, **7(6)**:R953-964.
53. Affymetrix: **Microarray Suite User Guide, Version 5.** Affymetrix 2001http://www.affymetrix.com/support/technical/manuals.affx.
54. Fan W, Khalid N, Hallahan AR, Olson JM, Zhao LP: **A statistical method for predicting splice variants between two groups of samples using GeneChip expression array data.** *Theor Biol Med Model* 2006, **3**:19.
55. Hu GK, Madore SJ, Moldover B, Jatkoe T, Balaban D, Thomas J, Wang Y: **Predicting splice variant from DNA chip expression data.** *Genome Res* 2001, **11(7)**:1237-1245.
56. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Res* 2003, **13(9)**:2129-2141.
57. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B: **Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools.** *Nucleic Acids Res* 2006, , **34 Web Server:** W645-650.

doi:10.1186/1471-2105-11-108
Cite this article as: Robinson *et al.*: SplicerAV: a tool for mining microarray expression data for changes in RNA processing. *BMC Bioinformatics* 2010 **11**:108.