IDENTIFICATION OF ALTERNATIVE EXON USAGE IN CANCER SURVIVAL USING HIERARCHICAL MODELING

BY:

AHMED SADEQUE

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Animal Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Advisor:

Professor Sandra Rodriguez-Zas

ABSTRACT

Background

Alternative exon usage (AEU) is an important component of gene expression regulation. Exon expression platforms allow the detection of associations between AEU and phenotypes such as cancer. Numerous studies have identified associations between gene expression and the brain cancer glioblastoma multiforme (GBM). The few consistent gene expression biomarkers of GBM that have been reported may be due to the limited consideration of AEU and the analytical approaches used. The objectives of this study were to develop a model that accounts for the variations in expression present between the exons within a gene and to identify AEU biomarkers of GBM survival.

Methods

The expression of exons corresponding to 25,403 genes was related to the survival of 250 individuals diagnosed with GBM in a training data set. Genes exhibiting AEU in the training data set were confirmed in an independent validation data set of 78 patients. A hierarchical model allows the consideration of covariation between exons within a gene and of the effect of the epidemiological characteristics of the patients was developed to identify associations between exon expression and patient survival. The same model serves multi-exon models with and without AEU and single-exon models.

Results

AEU associated with GBM survival was identified on 2477 genes (P-value < 5.0E-04 (FDR adjusted P-value < 5.0E-04). G-protein coupled receptor 98 (*Gpr98*) and epidermal growth factor (*Egf*) were among the genes exhibiting AEU with 30 and 9 exons associated with GBM survival,

respectively. Pathways enriched among the AEU genes included focal adhesion, ECM-receptor interaction, ABC transporters and pathways in cancer. In addition, 24 multi-exon genes without AEU and 8 single-exon genes were associated with GBM survival (P-value < 0.0005).

Conclusions

The inferred patterns of AEU were consistent with *in silico* AS models. The hierarchical model used offered a flexible and simple way to interpret and identify associations between survival that accommodates multi-exon genes with or without AEU and single exon genes.

ACKNOWLEDGEMENT

All praise to Almighty **ALLAH (S.W.T)**, the omnipotent, the most compassionate and His prophet **Muhammad (P.B.U.H),** most perfect amongst those born on earth, who is a beacon of guidance and knowledge for humanity as a whole. I wish to express my sincere gratitude to everyone who has contributed to this thesis and helped me along the way.

First and the foremost, thanks to my worthy advisor **Prof. Dr. Sandra Rodriguez-Zas**, Department of Animal Sciences, University of Illinois at Urbana-Champaign, USA for giving me the opportunity to work under her supervision. It was because of her inspiring guidance and dynamic supervision during the entire study program that I could complete this manuscript.

I am immensely grateful to my **PARENTS** and my brothers (**Ahsan** and **Ahwaz**) who have supported me throughout this time period and made me what I am today. I am also grateful to **Mr. Bruce Southey, Mr. Malik Nadeem Akhtar, Mr. Nicola Serão and Mr. Zeeshan Fazal** for sharing their expertise in Bioinformatics & Statistical Genomics and for their advice and guidance throughout this work. My sincerest gratitude to **Dr. Jonathan Beever** and **Dr. Juan J. Loor,** for being part of my thesis defense committee.

AHMED SADEQUE

TABLE OF CONTENTS

CHAPTER 1

LITERATURE REVIEW

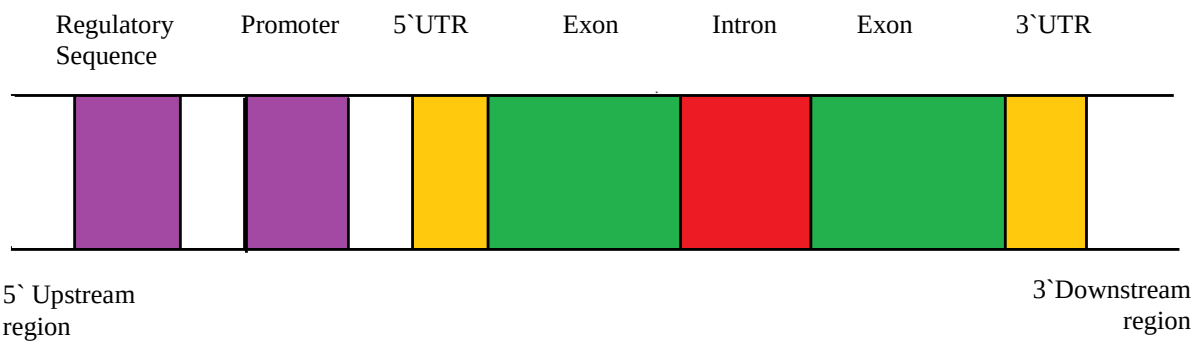
Alternative Splicing of Genes

Central Dogma and Alternative Splicing

Gene regulation is an important aspect for life and is a fundamental process in all living cells. Various fundamental mechanisms such as transcriptional factor regulation, transcriptional machinery, reproduction, homeostasis and adaptation all contribute in regulating gene expression. Studies have revealed mechanisms such as chromatin modifications, transcription, splicing, other mRNA modifications, translation and post translational modifications impact gene expression [1]. In 1958 Crick proposed the "Central Dogma of Molecular Biology", which enunciated the transfer of information between DNA, RNA and protein. According to central dogma the transfer of genetic information can be subdivided into three major categories: transfer with direct or indirect evidence (DNA→DNA, DNA→RNA and RNA→Protein), transfer with no evidence (RNA→DNA and DNA→Protein) and no transfer (Protein→Protein, Protein→DNA and Protein→RNA), where '→' represent the flow of information. Hence, according to central dogma genes are transcribed into RNA molecule and further translated into a polypeptide chain [2]. Another important concept "one gene one enzyme" was put forward by Beadle and Tatum in 1941, which in 1962 was modified to "one gene one polypeptide" by Ingram. The completion of Human Genome Project (HGP) lead scientists to establish that human genome consists of roughly 30,000 genes, where the genes present in the genome are of varying lengths. Gene can be defined as, "the complete sequence region necessary for generating a functional product", including both protein-coding and non-coding RNA genes. Most of these genes consist of coding region that are expressed referred to as 'exon' and non-expressing

intervening sequences known as ‘intron’ (Figure 1.1) [3]. In 1977, Walter Gilbert became the first one to suggest the concept of exon and intron and suggested that different mRNA variant can be produced from the same gene by splicing various exonic combinations. Walter observed that genes of eukaryotes contained intervening sequences that were removed as post-transcriptional modification and were referred to as ‘Introns. By 1980’s various studies recognized that Alternative Splicing (AS) as a natural process occurring in the genome by confirming the presence of different transcripts of the same gene. Based on the number of expressed sequence (mRNA), it was anticipated that humans would have a much larger genome than drosophila (14,000 genes), including approximately 150,000 genes. Estimates from different studies suggest that approximately 95% of all the human genes are subjected to AS [4]. However, the sequencing of human genome reported presence of some 32,000 genes which were far less than anticipated. This vast difference in human gene content led scientists to evaluate the importance AS in producing genomic variation [5]. Thus, an AS event is categorized by the formation of different isoforms, mRNAs with altered gene functions produced from the same locus possessing different protein coding DNA sequences (CDS), transcription start sites (TSS) and untranslated region (UTR), from the same transcript due to retaining different exonic segments and splicing different combinations of splice site together in the mRNA [6]. Thus, the phenomenon of AS is a vital cellular and regulatory process involved in regulating genes, as variety of processes ranging from cell growth and differentiation to apoptosis utilize AS for their proper functioning, and diversifying genome by compelling genes with multiple exons to produce distinct variants that in turn code for structurally and functionally distinct protein variants i.e. involved in regulating and generating genomic and proteomic diversity [7].

Figure 1.1: Structure of a typical gene



Alternative splicing can be divided into three broad categories; intron retention, cryptic splice site usage (functions by elongating or shortening the exon), and alternative exon usage (AEU). An AEU results in the skipping of exon and is also termed as alternative 5` to 3` splicing. Alternative exon usage is sub-divided into two categories: cassette exons (discrete exons that can be independently included or excluded) and mutually exclusive splicing (which involves the selection of only one from a group of two or more exon variants [4].

Mechanism For Regulating Alternative Splicing

A variety of genetic diseases manifest solely due to mutation in splice site sequences, spliceosome complex and auxiliary or cis-regulatory elements including: Exon or Intron Splicing Enhancers (ESE and ISE) and Exon or Intron Splicing Silencers (ESS and ISS) [6]. The exclusion or inclusion of genomic content in a transcript is governed by a ribonucleoprotein complex called Spliceosome through the process of splicing [6, 8]. Splice sites are present at each exon intron boundary, where during splicing introns are removed from pre-mRNA and the exons are then spliced together.

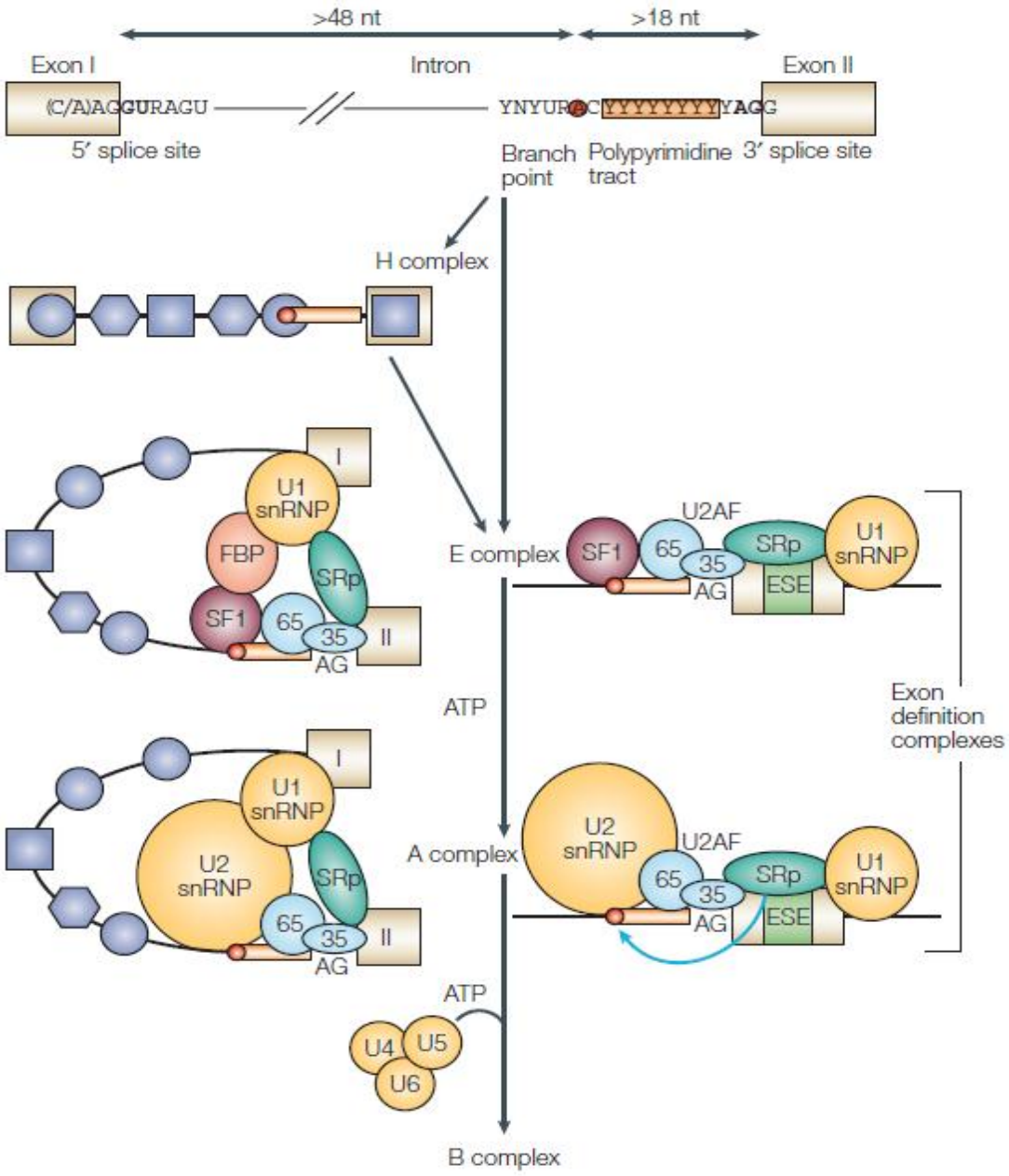
Both spliceosome complexes composed of five small ribonucleoprotein particles (snRNPs) and auxiliary proteins are important components of the splicing machinery. The auxiliary elements and spliceosome complex function collectively to correctly identify the splice site. The initial stage of spliceosome complex assembly involves recognition of the 5` splice site by snRNP U1 followed by binding of splicing factor 1 (SF1) to the branch point and binding of U2 auxiliary factor (U2AF) to both polypyrimidine tract and 3` terminal AG. This assembly of the complex is referred to as E-complex. The ribonucleoprotein U2 snRNP then binds to the branch point replacing the SF1and converting the complex to into an ATP-dependent complex: pre-

spliceosomal A-complex. This complex is then converted to B-complex following the recruitment of U4/U6-U5 tri-snRNP. Conformational changes in the structure of B complex leads to the formation of catalytically active spliceosome "C complex" (Figure 1.2). Various mechanisms at different steps of spliceosome assembly can contribute in regulating an AS event. Few steps through which AS can be regulated are correct identification and selection of the splice site, U1 & U2 snRNP base pairing, transcription coupled AS, tissue-specific AS [9].

Inhibiting Splice Site Recognition

U1 snRNP are recruited to the 5` splice site, similarly U2AF complex and snRNP are also recruited towards the 3` splice site by the SR protein after binding with ESE. Domains containing Arg-Ser repeats (RS) mediate the interaction between SR proteins, ESE and protein targets. On the other hand, there are many means to evade the recognition of the splice site. For instance splicing silencers can block the access of snRNPs or other positive regulators, if they are present in close proximity of the splice sites or splicing enhancers. One such example is of polypyrimidine-tract binding protein (PTB) that inhibits the binding of U2AF to regulated exon by binding to polypyrimidine tract [10]. Activators are also blocked from binding to enhancers by the splicing inhibitors. Inhibitors such as FOX1 and FOX2 prevent the recognition of splice site by binding to exonic sequence, in close proximity of ESE, in pre-mRNA. This binding inhibits binding of other activators (TRA2 and SRp55) to the ESE, preventing the U2AF recruitment by the activators, thus inhibiting the formation of E-complex [11].

Figure 1.2: Basic principle of Splicing[1]



[1]Reprinted by permission from Macmillan Publishers Ltd: NATURE REVIEWS | MOLECULAR CELL BIOLOGY [6], copyright (2005).

Splicing silencers, also inhibit splice site and reside 100-200 b.p. away from the enhancers. Such silencers are considered to function by assembling multimers along RNA, thereby causing splice site to become inaccessible [12]. It is also proposed that the protein-protein interaction between RNA-binding proteins residing in close proximity of alternative exon, causes the alternative exons to loop out, which inturn poses hindrances in further spliceosome assembly [13].

Protein Factors Regulating Alternative Splicing

A functional spliceosome results from cross interaction between introns of U1 and U2 snRNPs. However, binding of hnRNP L to ESS can potentially inhibit the pairing of U1 and U2 snRNPs15. An ATP-dependent spliceosome-like complex, known as A-like exon-definition complex (AEC) is very similar to A-complex and contains both U1 and U2 snRNP. The progression of this complex into the B complex is inhibited by hnRNP L-containing AEC in presence of hnRNP L, which functions by inhibiting the cross-intron complex formation between adjacent U1 and U2 snRNPs bound to exonic splice sites, resulting in exon-skipping. On the other hand, U4/U6–U5 tri-snRNP complex is recruited in absence of hnRNP L and after formation of A-like complex across exons, converting intron defined A complex to form the B complex. It is possible that hnRNP L hinders the formation of B-complex by either substantially shielding the interaction between the two snRNP or by introducing conformational changes in the mRNA. This change renders the mRNA unable to form cross-intron pairing between snRNP-bound alternative exons.

Though most of the studies have related inhibition of intron definition with splicing regulation, considering occurrence of AS due to activation of intron definition cannot be left out of the equation. A study conducted by Martinez-Contreras et al. (2006) [14] resulted in verification of

presence of hnRNP binding sites in close proximity of intron definition sites assist in splicing. Presence of hnRNP sites close to each other results in cross-intron interaction between hnRNP's. This interaction causes the intron ends to come closer to each other. Thus, indicating that hnRNPs are indeed one of the probable explanations that promote splicing [15]. Proteins such as cis-acting elements also play significant role in splicing regulation, certain cis-acting elements control the binding at 5` splice site. Study conducted by Yu et al. (2008) [16] revealed presence of ESSs and ISSs that inhibit the recognition of proximal (strong) 5`splice site by selecting a weaker (distal) 5` splice site, thereby altering U1 and U2 snRNP pairing. The inability of splice factors to bind to the correct splice site is due to occurrence of conformational change in proximal 5` splice site complex by action of splicing silencers. As a result of which proximal splice site complex loses its edge over the distal splice site complex for binding to U2 snRNP 3` splice site complex [15].

Tissue Specific Alternative Splicing

Numerous splice variants are produced from multi-exon genes subjected to AS. This AS also results in production of structurally and functionally distinct protein products. Understanding the principles governing the splicing differences has become a necessity. Even more alluring fact is that studies have shown presence of different isoforms from the same gene in different tissues suggesting involvement of different factors. Microarray techniques have played a huge role in determining tissue-specific regulation of AS events but can only provide us with a limited amount of information due to the inability to distinguish between closely related variants. In tissues, factors typical to tissues and regular RNA-binding factors act in combination, influencing spliceosome to regulate AS event. RNA map, provides a good understanding on the

effect of various factors on AS, based on the site of action [15, 17]. Expression of splice factors specific to tissues and regulation of mRNA's (targeted by the splice factors) can contribute towards the understanding of tissue-specific AS [15]. Wang et al. (2008) [17] verified the differences in AS occurring in different tissues by taking into account ~105,000 AS events of 8 different types. The results deduced from the analysis showed disparity between tissues in amount of splice variants produced. The isoforms detected to be differentially expressed between different tissues were also taken into account. The frequency of identified tissue-specific AS event through the analysis was approximated to be over 22,000. The conclusion deduced from the experiment indicated that most of the AS events are subjected to tissue-specific regulation [17].

Brain usually displays the most frequent number of splicing patterns, with several identified splicing regulatory factors. One of the splicing regulatory factors is nPTB, which is highly expressed in differentiated neurons. However, neural progenitor cells show high expression of PTB. Thus, presence of PTB-to-nPTB switch promotes the post transcriptional mechanism necessary for differentiating the neuronal cells.

Impact Of Activators And Inhibitors

Both splicing activators and inhibitors pose huge impact on the splicing procedures; a pre-mRNA undergoes. The combinatorial, as well as the competitive effects of activators and inhibitors effect the inclusion or exclusion of an alternative exon in the transcript. However, it is the activity of particular regulator type that regulates the fate of alternative exon [18]**.** A study conducted by Crawford and Patton (2006) confirmed that different regulatory factors exclude or include an exon by competing to bind to the same element. It was shown that the competitive

struggle between SR proteins and hnRNPs for binding to the same element regulates the exon-2 of α-tropomyosin [19].

Position Specific Splicing

Different studies have concluded that occasionally it is the positioning of the auxiliary elements (ESE, ISE, ESS and ISS) and their binding proteins that regulate the exon. Depending on the location of the binding site relative to the exons, the proteins can function as both activators and repressors [20, 21]. NOVA1 binds to ESS residing in the exon-4 of its own mRNA, thus excluding the exon-4 [22]. On the other hand, the inclusion of exon-9 in pre-mRNA of GABA A receptor γ2 (*Gabrg2*) gene is promoted by binding of NOVA1 to ISE [23]. A study conducted by Ule J et al (2006) [24] focused on generating mRNA maps that contained locations for the binding sites of NOVA1 and NOVA2 and the significance of binding. The maps were developed by searching for the existence of YCAY clusters that are the binding sites of NOVA1 and NOVA2, and by utilizing prior information from mRNA's targeted by the NOVA1 and NOVA2 [24]. Hence, results clearly indicate that the positioning of the splicing regulatory elements and proteins with respect to the regulated exon is an important aspect that cannot be neglected while considering AS. Binding of factors responsible for splicing to enhancers or silencers brings forth structural changes in the mRNA, presenting or obstructing the access of splicing machinery to the splice sites of alternative exons [15].

Involvement Of RNA In Alternative Splicing Regulation

Secondary structures of pre-mRNA pose a substantial influence over the selection of splice site. It has been established that the secondary structures can affect splice sites by either binding competitively to splice factor binding sites or by masking the splice sites. *Dscam* mRNA in drosophila presents a complex alternative splicing event. Down syndrome cell adhesion molecule (*Dscam*) exon-6 cluster comprises of 48 mutually exclusive exons, in which pairing of conserved sequences, downstream of exon-5 (dock site) & upstream of every exon in exon-6 array (termed as 'selector' sequence) results in inclusion of one exon-6 variant in the transcript. Binding of hrp36 (selector sequence homolog) leads to exclusion of other exon-6 variants [15, 25].

Riboswitches are part of mRNA molecules which effect gene activity by binding to the target genes. It is well established that riboswitches control gene expression in prokaryotes [26] and are involved in regulation of AS in some smaller eukaryotes. However, their involvement in AS event of higher organisms remains to be established. [27]. Small nucleolar RNAs (snoRNAs) have been associated with AS event regulation. It was observed that AS is accomplished when snoRNA HBII 52 binds to the silencing element in exon Vb of serotonin receptor (5-HT_{2C}R) to promote its inclusion [28].

Prediction Of Alternative Splicing

Various bioinformatics approaches have been developed to predict AS events. Su et al. (2008) [29] developed an individual exon approach and Purdom et al. (2008) [30] used the residuals from probe level analysis to identify AEU on a per-sample level in the FIRMAGene package. In this approach, series of consecutive residuals that depart from 0 are considered evidence of AEU

events. The sample-level analysis challenges the detection of AEU events or the identification of common patterns across treatments or conditions. The limitations of this approach were overcome by the use of a linear model to compare the exon expression among groups (Laderas et al., 2011) [4]. However, group comparison is not suited to identify alternative exon usage associated with other conditions such as survival or time-to-event. In addition, the previous approach did not account for correlations between exons measured on the same sample. Cline et al. [31] formulated a ANalysis Of Splice VAriation ANOSVA approach that models the logarithm of the background corrected intensity *(yij)* as a function of two factors: concentration θj of the target mRNA in the sample (target concentration) and a probe affinity term φi relating changes in probe intensity to unit changes in target concentration ($yij = \varphi i + \theta j +$ error). However, a limitation is that the model is not applicable to a gene that produced more than one splice form [31]. Barash et al. (2010) [8] developed a probabilistic model that was able to identify AS signals specific to a particular condition by utilizing prior knowledge about the AS event such as (information corresponding to expression level correlation, effect of AS event and expected small number of AS signal) and dataset based specific knowledge including (levels of gene expression in dataset and measurement quality pertaining to the dataset) [8]. Shai et al 2006 [32] developed a model namely Generative model for the Alternative Splicing Array (GenASAP) that predicts exon skipping event by quantifying splicing changes at single exon level. AS levels in microarray data are predicted by GenASAP, which utilizes Bayesian learning in an unsupervised probability model [32]. Studies conducted by Eisen et al. (1998) [33] applied clustering analysis to group genes by detecting similarities in patterns of gene expression. The same technique can be applied for identifying common patterns in AS events. A matrix representing inclusion and exclusion isoform for each exon against specific condition can help in

identifying patterns pertaining to AS events [33, 34]. However there are many drawbacks to this technique, firstly it cannot account for tissue specific splicing. Also the clustering either distinguishes or neglects to distinguish between relative increase or decrease due to exon inclusion or exclusion [8]. Zheng et al. (2009) developed a Regression Method for AS detection (REMAS) that was based on the Lasso regression algorithm [35]. This approach assumes that the overall gene expression is strongly interrelated with the intensities of constitutive exons and an AS event results in altered gene expression. Thus, the difference between the expression of alternatively spliced and constitutive exon is considered as an indicator of AS. Finally, the Splicing Index (SI), a basic linear model for estimating changes of exon expression, applied in this model, ignores the relation between exons and identifies alternatively spliced exons individually [35].

A number of resources and databases have also been developed to provide and store information for correct prediction of AS event. Dralyuk et al. (2000) [36] developed Alternative splicing database (ASDB) using Genbank and SWISS-PROT annotation. This search engine allows queries to be searched across SWISS-PROT and GenBank fields and then simply following the links to all variants allows information regarding splicing event to be retrieved [36]. In 2001, an alternative splice database of mammals (AsMamDB) was developed by Hongkai et al. (2001) [37] to assist studies related to alternatively spliced genes of mammal. In AsMamDB alternatively spliced genes are associated with a cluster of nucleotide sequences. The main information provided by AsMamDB includes AS patterns, gene structures and also provides information about gene products and gene's expression site [37]. Intron information system (ISIS) was developed to evaluate the function and evolution of spliceosomal introns in eukaryotes. Analysis through this system allowed recognition of many alternative spliced exons

[38]. HASDB, a database to detect AS events in human EST data was established by Modrek et al (2001) [39]. The results obtained through HASDB provide deep understanding of AS function in human genome [39]. Another bioinformatics resource for AS events ASD was initially developed by Thanaraj et al. and was upgraded by Stamm et al. (2006) and contains both manually and computationally generated data[40]. This resource functions by collecting and annotating data related to AS. This resource consists of various parts: AltSplice, AEdb and a Workbench. AltSplice is a database that includes computationally predicted alternatively spliced events, patterns and transcripts. Gene alignments are utilized by AltSplice to generate the data. Information on various features including splicing signals, SNP-mediated splicing, intra-specie homology and expression states is provided by AltSplice. Results obtained from this component indicated that about 61% of human genes undergo AS. It was also concluded that approx. 3.9 alternatively spliced transcripts are produced form a single gene. Around 5200 orthologous gene pairs (between human and mouse) are included in AltSplice. AEdb is manually developed portion of ASD, it contains datasets that are based entirely on literature. It can further be divided into 4 components: *AEdb-Sequence* (searches pub-med for studies related to AS), *AEdb-Function* (provides with literature-based survey of functions related to a particular alternatively spliced exon. The functionalities of proteins generated as a result of AS event are divided into 11 different categories), *AEdb-Motif* (provides with literature related to splicing regulatory motifs, intronic/exonic regulatory sequences and mutations. It reported some 153 and 81 enhancers and silencers respectively) and *AEdb-Minigenes* (provides graphical representation of splicing patters and regulatory sequences for all minigenes reported in literature and includes a collection of 82 minigenes). The last component is a workbench that is used for analysis of splicing. This system allows retrieval of information on variety of aspects including intron characterization across

splicing signals, identification of splicing regulatory elements, prediction of putative exons and translation start codons [40]. Another database developed that accumulates lot of information regarding AS events is The Alternative Splicing and Transcript Diversity database (ASTD). ASTD comprises of vast collection of alternative transcripts that integrate transcription initiation, polyadenylation and splicing variant data. Alternative transcripts are derived from the mapping of transcribed sequences to the complete human, mouse and rat genomes using an extension of the computational pipeline developed for the ASD (Alternative Splicing Database) and ATD (Alternative Transcript Diversity) databases, which are now superseded by ASTD. ASTD datasets are established through three different categories of transcript-to-genome mapping. The three prediction categories include splicing (AltSplice), polyadenylation (AltTrans and AltPAS) and transcriptional start site (AltTSS) variant. Altsplice was used for predicting spliced isoforms and AS events by mapping EST and mRNA onto genomic sequences of Ensembl. ASTD contained 8,125,884 mapped transcripts for humans, 4,935,071 for mouse and 824,394 for rats. However, after removing all sorts of false positive transcripts, less than 25% of the true positive mapped transcripts remain in ASTD that can be used as supports for splice variants. Two components are involved in the polyadenylation: AltTrans and AltPAS, both are responsible for identifying polyadenylation but recognize sites corresponding to specific splice patterns and potential poly (A) sites irrespective of underlying splice patterns, respectively. Splice sites were included only if they met the specified criterion for internal priming, unmatched transcript ends and presence of polyadenylation signal that has already been reported. The Poly (A) sites obtained from both the components are then merged. The last component AltTSS was used for predicting transcription start site (TSS) using libraries of oligo-capped full-length cDNA's. TSS were taken into consideration after aligning the sequences residing ~10,000 b.p. upstream of

each splice variant with NCBI Blast program. High-scoring sement pairs with minimum 95% identity were selected and onwards filtered with the specified criterion. Results provided by ASTD for human genome report 68% splicing variants, 68% transcription initiation variants and 62% polyadenylation variants.

Annotating genes, transcripts and proteins is a complex and difficult task. Tools that can correctly predict genes have become a requirement. AceView predicts gene models and provides with non-redundant and comprehensive graphical and sequence representation all public mRNAs by summarizing cDNA data from Refseq, GenBank and dbEST. Developed by NCBI, AceView utilizes heuristics for maintaining the same annotations. AceView is also displayed as one of UCSC gene tracks. Through analysis of the gene prediction tools, it was established that AceView transcripts are more close to Gencode as compared to other transcripts predicted by other tools. Therefore, for all transcripts AceView annotates the finest predicted CDS. Introns of both Gencode and AceView are common (except 10% and 14% specific to Gencode and AceView respectively), nucleotides used in spliced variants are common (except 8% and 12% specific to Gencode and AceView respectively). AceView also provides with a more efficient and simplistic method for annotating complete chromosome, while maintaining a similar annotation quality as to Gencode. One other feature provided by AceView is re-annotation of its mRNA with parsimonious Gencode-like CDS. Overall results indicated that gene structures predicted by AceView are in agreement with those of Gencode [41].

Alternative Splicing Influences Health

Alternative splicing has been associated with a variety of cellular, molecular and physiological functions. The regulation of these mechanisms by AS displays even more variability than their regulation through promoters [42]. Studies have concluded that AS is the causal agent of variety of diseases ranging from developmental regulation, cancer to apoptosis [43]. Numerous novel protein products with completely different peptide sequence, structure and functions are produced as a consequence of AS [43]. Alternative splicing affects most of the genes residing in the genome, and thus changes occurring during the transcriptional and translational processes might manifest in disease. The aberrant pre-mRNA processing might be instigated because of mutations in cis elements or altered expression of splicing factors and can potentially lead to tumoral transformations and cancer development [44]. Kim et al. established that aberrant mRNA and the resulting proteins have distinctive characteristics and properties that impart distinctive growth, differentiation and other molecular properties to the cancerous cells [45].

Alternative Splicing Necessary For Developmental Processes

Occurrence of alternatively spliced events for generating genomic and proteomic diversity has been related to the proper functioning of many biological processes. Grabowski et al., (2001) [46] declared that AS was the primary cause of protein diversity required for proper functioning and development of the Nervous System (NS). The study suggest that AS of exon 21 residing in N-methyl-D-aspartate R1 (*NmdaR1*) receptor is responsible for many important regulatory processes in brain like neuronal development and synaptic plasticity. Other studies [47] conducted in 1995 concluded that C1 cassette exon containing *NmdaR1* receptor mRNA can also be directed to function in plasma membrane. The mRNA isoform expression was inspected in

quail Qt6 fibroblasts cell line. The receptors were clustered on plasma membrane only when the gene contained C1 cassette exon. In absence of this cassette exon, the target protein was not observed in plasma membrane [46]. It has been established that C1 exon is required in gene to associate *NmdaR1* receptor with the neurofilaments [48].

Wu et al. (2010) [49] demonstrated the importance of AS by associating it with complex biological system such as cellular apoptosis. Apoptotic pathway is initiated after interaction between specialized TNF family ligands with their receptors. Extracellular domains are proteolytically cleaved to generate soluble form of liquids [50]. As shown by Agarwal., 2003 FasL variants are soluble and pose a significant influence upon apoptotic potential by blocking the death-promoting activity. These soluble isoforms of FasL, which are deficient of intracellular domain, transmembrane and portion of extracellular domains inhibit apoptosis and are generated due to AS event [51]. Another important gene is the *Bcl-2* family, many members of this family have been associated with apoptosis inducing and inhibiting activity. Study concluded by Adam., 2003 suggested that AS is involved in regulation of many *Bcl-2* proteins. *Bcl-x* is a member of *Bcl-2* family and is subjected to AS, producing two functionally separate isoforms: *Bcl-*x_L and *Bcl-*x_S. *Bcl-*x_L is the longer transcript comprising of all four BH domains and functions by inhibiting apoptosis. On the other hand, produced as a result of AS, Bcl-x_S is the smaller of the two transcript lacking both BH1 and BH2 domains. As opposed to the expression of *Bcl-*x_L in long lived cells, expression of Bcl-x_S is normally observed in cells enduring high turnover rate and in hormone-dependent tissues [50].

Alternative Splicing causes disease

Different studies have displayed abnormal AS to be related with a variety of diseases. One of the diseases that manifests due to AS is the spinal muscular atrophy (SMA) that is characterized by the degeneration of alpha-motor neurons in brainstem and spinal cord. It affects approx. 1:10000 infants world-wide. In most cases manifestation of the disease in infants leads to death in early childhood [52]. Aberrant assembly of snRNP is reported to cause SMA, this abnormal assembly occurs due to loss of *Smn1* gene which is responsible for producing SMN protein [43]. The study conducted by Zhang et al. (2008) [53] showed that motor neurons remain the only ones that are affected by splicing, no other defects leading to cell death were observed due to splicing. Exon arrays were utilized to compare splicing difference between 3 normal and 3 *Smn1*deficient mice. The analysis concluded 259, 73 and 633 from spinal cord, brain and kidney respectively, to be significant at a false discovery rate (FDR) adjusted P-value < 0.1, while utilizing 200,000 probes that corresponded to exons of some 20,000 mouse genes. To confirm the results obtained, exon-junction specific primers were designed to conduct real time RT-PCR on 31 genes that were displayed significant by the exon arrays. The results obtained from RT-PCR validated the exon array results with a rate of 97%, suggesting differential expression of the exons in a particular tissue. They also used 8 genes to confirm the expression of same exons across tissues. These findings revealed that different tissues possessed disparity in expression levels of the exons, indicating that the splicing alterations are tissue specific [53].

Familial Hypercholesterolemia (FH) is a metabolic disorder characterized by the presence of elevated levels of total cholesterol (TC) and low density lipoproteincholesterol (LDL-C) affecting cholesterol metabolism. In some cases patients suffering with FH display skin and tendon xanthomas, where FH manifests itself into premature coronary heart disease (CHD).

Mutations in three genes have been shown to cause hypercholesterolemia (HC) these are: Low Density Lipoprotein Receptor (*Ldlr*), Apolipoprotein-B (*ApoB*) and Pro-Protein Convertase Subtilisin like Kexin Type 9 (*Pcsk9*) [54]. However, primary cause of manifestation of HC is mutation in *Ldlr*. Zhu et al. (2007) [55] analyzed SNPs present in *Ldlr* relative to ESE matrices and discovered presence of C/T (ESE site) single nucleotide polymorphism (SNP) rs688 in exon-12 of *Ldlr* that enhances splicing event resulting in exclusion of exon-12 from the transcript. *Ldlr* c-DNA from exon 10-14, from liver samples of 21 female and 22 male patients, was amplified to analyze the splicing event. *Ldlr* isoforms missing exon-12 were readily detected and genotyping of rs688 revealed presence of T allele (minor allele) and its presence decreases the splicing regulatory protein (*Spr40*) binding affinity, which recruits the splicing machinery, especially in pre-menopausal women (P-value<0.0042). Splicing pattern caused by SNP is significantly related to high cholesterol level in women only (P-value<0.024) [55]. A truncated receptor is generated as a result of Exon-12 skipping, this receptor is deficient of transmembrane domain. Thus, no internalization and membrane binding occurs preventing LDL uptake by the cell. LDLR also acts as a receptor for Apo lipoprotein E (*ApoE*), which is reported to be associated with development of Alzheimer's disease. Zou *et al*. (2008) [56] concluded that skipping of exon-12 in Ldlr transcripts increases the possibilities of occurrence of the disease in the male, while no association of splicing event with alzheimers was observed for women [43]. This tissue-specific AS can be held accountable for the disordered cell differentiation and signaling that contribute to stem cell like proliferation of cancer cells [57] .

Alternative Splicing Associated With Cancer

Alternative splicing leads to formation and expression of numerous different transcripts, produced due to varying combinations of exon inclusion and/or exclusion. The translation of these transcripts may results in production of structurally different protein that also possess different functions [57]. Numerous studies have validated the presence of alternative splicing patterns in cancerous cells. Studies have also associated aberrant AS event with development of cancer. The variable expression of these AS or tumor-specific spliced variants triggers many cellular and molecular functions that promote proliferation, motility and division of cancerous cells [43]. Cancerous cells have been reported to disrupt the splicing patterns by two prominent methods involving somatic mutation in cis-elements and trans-acting factors involved in regulating splicing [58]. Mutations in trans-acting factors, on numerous occasions have been been reported to be associated with various cancers including glioma, ovarian and colon cancer. Additionally, many gene have been reported with normal splice variant that contributes towards the development of tumorigenesis [57]. However, the roles of spliced variants in cancer have not been fully established; presence of a spliced variant in malignant phenotype could be a coincidental i.e. the spliced variant could be present in the malignant phenotype without ever contributing to its development [43]. However, the fact remains that results obtained from both predicted and experimentally verified data claim that AS is more prevalent in cancerous cells. On the other hand, study conducted by Kim *et al*. (2008) confirmed presence of relatively lower degree of alternatively spliced exons in cancerous cells as compared to normal cells [45].

Breast cancer susceptibility gene (*Brca1*) has known association with development of hereditary breast and ovarian cancers and is also reported to produce several splice variant that might significantly contribute to the development of tumor. Occurrence of mutations in intronic splice

sites and degenerative sites, located near intron/exon boundary, result in development of numerous splice variants. One such splice variant, resulting in exclusion of the constitutive exon-18, involves G>T mutation at position 6 of exon-18, leading to E1694X change and removal of 26 amino acids (a.a).The mutation disrupts the C-terminus of BRCA1. It was hypothesized that the mutation in consideration disrupted the ESE. However, the mutation occurred in a region not rich in purines whereas ESE's are normally expected to be residing in purine-rich region. Utilizing motif scoring matrices it was established that the mutation disrupts ESE due to confirmation of correlation between SF2/ASF high-score motif distribution and the splicing patterns. Thus, SF2/ASF recognition sequence is a necessary; absence of recognition sequence can lead to skipping of exon-18 [59].

Glioblastoma Multiforme

Background

Appropriate neuronal cell differentiation is necessary for the proper functioning of brain cells. Misregulation in neurotransmitter signaling, sequence mutations, methylation patterns, copy number variation, faulty apoptosis, erroneous DNA repair and cell differentiation can result in stem cell like proliferation of cells leading to development of brain malignancies [57].

The World Health Organization (WHO) has grouped brain cancers into four different groups (I, II, III and IV) based on the severity of the disease, where group IV represents the most malignant tumors. Glioblastoma Multiforme (GBM) is considered as one of the primary and highly aggressive brain tumors, accounting for 50% of all CNS malignancies, 20% intracranial tumors [60] and 90% of all glioblastoma. GBM usually forms in cerebral white matter, exhibits devastating consequences with average survival ranging up to approximately 12 months and has been placed in group IV by WHO, due to high capacity of GBM to proliferate in the brain [61]. Secondary GBM referred to as 'Astrocytoma', display slow progression, accounts for less than 10% of all GBM cases and occurs in relatively younger patient.

The average survival time for a patient suffering with glioblastoma was estimated to be around five months. Increased time period after manifestation of disease resulted in decreased survival rates (42.4% at 6 months, 17.7% at 12 months and 3.3% at 24 months). It was also determined that males are slightly more susceptible to primary glioblatoma than females with male to female ratio being 1.28 [62]. However, females are more susceptible to secondary glioblastoma. Furthermore, age of an individual also poses significant effect over survival. Older age at diagnosis reduces the survival rates, while patients with age less than 50 years showed significantly longer survival and show higher incidence of rare secondary glioblastoma.

Manifestation of of primary tumors is observed in majority of older patients diagnosed with GBM [63]. Johnson et al. (2011) concluded that treatment of GBM patients with radiation and drug therapy increases the survival from 12 to 15.6 months [64]. However, even with all the therapies provided, GBM still displays high resistance to treatment because of presence of small areas displaying necrosis and hemorrhage in the tissues [65]. Studies also confirmed that ethnicity also contribute significantly in GBM development. It was estimated that white people are more likely to develop brain tumors as compared to non-white people [66].

Genes And Gene Expression Associated With Glioblastoma Multiforme

In most of the GBM cases, aberrant genomic alterations are associated with the development of tumorous cells. Therefore, understanding the involvement of underlying genes, pathways, mechanisms and functions that contribute towards the development of brain malignancies is a must [65]. Various genes including *Egf, Nf1, Idh1* have been associated with the initiation and progression of GBM [67].

Transcription is promoted by activation of several tyrosine kinases and downstream signal molecules after binding between epidermal growth factor (*Egf*) and epidermal growth factor receptor (*Egfr*). This binding also results in dimerization of Erg receptor family (Erb 1-4) [68]. Studies confirmed association of polymorphism, 61 A/G, with poor survival in GBM [69]. Another study performed tagging of *Egf* to estimate effect on the development of GBM. Results indicated minor alleles of four polymorphic events, rs17238095, rs3796944, rs9992755 and rs11568994 located in different exons, to be significantly associated with the GBM [68]. Therefore, analysis of various studies relates higher expression of Egf to be associated with GBM.

Neurofibrin1 (*Nf1*) is another gene that is negatively associated with GBM, by acting as a GBM suppressor gene. Accumulation of 19 specific mutations (6 non-sense, 4 splice site, 5 missense and 4 frame-shift indels) was observed and it was predicted that these mutations are responsible for the probable inactivate the gene. In all it was estimated that somatic mutation manifested in 23% of the total patients were contributing towards the inactivation of the gene and in the process confirming the significant association to GBM [65].

Isocitrate dehydrogenase 1 (*Idh1*) is a gene residing in chromosome-2 and has been associated with secondary glioblastoma [70]. IDH1 is responsible for the production of NADPH by acting as a catalyst in the formation of α-ketoglutarate from isocitrate through the process of oxidative carboxylation. Experiments conducted showed that mutations in *Idh1* reduced the activity of enzyme IDH1 due to formation of heterodimers that are catalytically inactive. This process also results in upregulation of a transcription factor hypoxia-inducible factor subunit (*Hif-1a*), which is reported to be associated with tumor growth [71]. The frequency of mutation of Idh1 was estimated to be above 80% in most of gliomas except primary GBM, which exhibits reduced frequency of less than 5% [70].

Genetic Pathways

Due to development of various techniques and methods, understanding of genetics underlying glioma development has greatly improved. Large scale sequencing of genome has led to detection of several novel genetic alterations and pathways, adding valuable information towards further understanding glioma and may also help in identification of targets for interventions [70].

Receptors (such as EGFR etc) normally reside in inactive state and get activated after binding to their respective ligands (EGF). Higher expression of Egfr is associated with development of primary GBM, where 70%-90% of GBM cases showing upregulation of *Egfr* also possess amplification of *Egfr* sequence. The interaction between EGFR and EGF results in recruitment of phosphatidylinositol 3-kinase (PI3K) complex composed of two subunits: catalytically active protein p110α and regulatory protein p85α. PI3K ultimately leads to activation of mammalian target of rapamycin (mTOR), a downstream effector molecule, by phosphorylation of phosphatidynositol-4,5-biphosphate (PIP2) to phosphatidynositol-4,5-3-phophate (PIP3). This entire process results in inhibition of apoptosis and consequentially promoting cell proliferation and survival. Phosphatase, tensin homologue, deleted on chromosome TEN (*Pten*) is a tumor suppressor gene that inhibits the cell proliferating action of PIP3. About 40% of the primary GBM cases exhibit mutation in *Pten*. The Cancer Genome Atlas (TCGA) pilot project reported alteration of Egfr/Ras/Nf1/Pten/Pi3k pathway in 88% of the GBM cases [15, 70].

Tumor protein p53 (Tp53) encodes for p53 protein that is reported to be associated with a variety of malignancies. Its main function is to regulate cells in response to increased cellular stress, cell death, differentiation. High levels of p53 protein are normally observed in malignant and transformed cells. It mainly consists of three domains including DNA-binding, transcription activation and oligomerization domains. The p53 protein is a DNA-binding protein which activates in response to DNA damage. Transcription of p21, Mdm2 gene is induced by activated *Tp53*. Amplification of *Mdm2* is associated with ~15% of glioma cases, binds to *Tp53*, thereby blocking transcriptional event induced by action of *Tp53*. Another gene *p14ARF* binds to *Mdm2*, blocking the Tp53 binding ability of *Mdm2*. Methylation in promoter region of *p14ARF* is observed in almost 50% of glioma cases. *Tp53* is also responsible for regulating *p14ARF* thereby

acting as a feedback mechanism. Another gene Mdm4 regulates the activity of Tp53. Thus mutation resulting in altered function of Tp53 can completely disrupt the pathways involving *Tp53, Mdm2, Mdm4* and *p14ARF*. Different studies have reported frequency of mutated Tp53 in primary glioma (65%) to be reduced than its presence in secondary glioma (28%). The Cancer Genome Atlas (TCGA) pilot project also reported alteration of *Tp53/Mdm2/ Mdm4/ p14ARF* pathway in 88% of the glioblastoma cases [15, 70].

Alternative Splicing Associated With Glioblastoma

Different studies to enumerate the expression of genes and their isoforms across tissues have been carried out. Ramskold *et al*., (2009) established that brain tissues contain expression of a large number of genes and gene isoforms because of high frequency of AS events [72]. AS is a natural process adopted by the genome to produce genetic and proteomic variation [4]. However, misregulation of AS has been associated with development of many diseases and cancers [59]. The development of GBM has also been linked with alternative splicing of various genes. One such gene is glioma-associated oncogene homologue 1 (*Gli1*), which is zinc-finger transcription factor. Gli1 protein functions as nuclear mediator for Hedgehog signaling pathway, a pathway known to regulate genes involved in premature development of the CNS and observed to be activated in gliomas. After being released from cytoplasm, Gli1 translocates to cell nucleus where its binding to GLI1-binding elements activates them. Lo et al., 2009 reported a presence of a truncated splice variant of *GLI1* (*tGLI1*). The new variant manifested in most GBM cells but was undetectable in normal brain cells. Further investigation of *tGLI1*, established that the variant is produced due to deletion of 123 bases from exon-3 and exon-4. The in-frame deleted portion contains 41 codons corresponding to specific amino acid residue position (34 to 74) in

the protein. Production of the variant *tGLI1* upregulates *CD24*, a gene reported to be associated with increased invasiveness. The results suggest that production of *tGLI1* results in gain-of-function that relates to aggressiveness of GBM cells due increased invasive and migrating properties of the infected cells [73].

Izaguirre et al. (2011) associated AS in (*Usp5*) with GBM. *Usp5* also referred as isopeptidase T is regulated by polypyrimidine tract-binding protein 1 (*Ptbp1*), whose upregulated expression level is responsible for cell proliferation and migration in GBM cells. PTBP1 protein is a splicing regulator and its members are responsible for repressing the recognition of exons during splicing, exon inclusion, replication, mRNA stability, RNA transport and viral translation. AS of *Usp5* leads to the formation of two isoforms: isoform-1 formed due to inclusion and isoform-2 containing exclusion of exons. This variation is observed as a result of 69 bases altering exon-15. Study revealed expression *Usp5* isoform-1 to be significantly correlated to *Ptbp1* expression. Results obtained were confirmed with RT-PCR and associated expression *Usp5* isoform-1 with reduced PTBP1 levels. *Invitro* studies also confirmed the presence of consensus PTBP-binding sequences in proximity of alternative exon. Thus, for isoform-1 presence of binding site specific to PTBP1 was observed at 5` splice site, resulting in exclusion of exon from isoform-1. Increase in the levels of Usp5 isoform was negatively correlated to GBM cells migration and proliferation [74].

Analysis Of Microarray Gene Expression

Understanding of genomic variations and differences in gene expression levels related to a specific phenotype requires knowledge of the underlying genes and pathways [75]. Transcription and translation of the gene into mRNA and protein respectively is an extremely complex and delicate process involving many regulatory factors [76]. Therefore, quantifying mRNA expression levels pertaining to specific phenotype is necessary for identifying its impact on the phenotype.

Gene Expression Measuring Platforms

Genetic makeup, environmental conditions, cellular response and regulatory factors all contribute to and are responsible for varying expression levels of the genes. These expressions levels are quantifiable through utilization of various techniques, particularly developed to measure gene expression over last two decades [77]. Most prominent methods used to quantify mRNA levels include Northern Blot [78], real time polymerase chain reaction (RT-PCR) [79], microarrays [80] and RNA-seq [81].

Northern blotting is a technique derived from southern blotting, in which enzyme-cleaved DNA fragments, separated due to movement of ions from positive to negative electrode are transferred to nitrocellulose strips. In northern blotting, single strands of DNA are coupled covalently to paper and are transferred through a gel. After which hybridization with labeled 32p probe is performed to correctly identify and detect specific sequence [78].

Microarray is a high throughput technique that was developed in the recent past and has the capability of measuring expression of thousands of gene in chorus. Microarrays utilize the

information obtained from mRNA and on assembly basis can be divided into two categories: spotted microarrays and oligonucleotide microarrays [80]. Through this methodology, manual arrays are developed that takes into consideration both research interests and cost limitation of the experiment. However, the results obtained through this technique are not very consistent [82]. On the other hand, oligonucleotide microarrays consist of arrays in which probes that are formed by adding each base individually. Additionally, much of the research being conducted globally nowadays utilizes oligonucleotide microarrays developed by different companies [80].

There are three main microarrays platforms that have been widely utilized for expression studies: two-dye approach, BeadChips and Affymetrix. The working of these microarray techniques is very identical. First step involves the hybridization of the labeled sample to the DNA probe, after the samples are applied onto the array. The complete process is affected by a variety of external and environmental factors including temperature, salt concentration, etc. The previous step is followed with washing of the array to ensure that hybridized targets are the only ones that remain attached onto the microarray. Thus washing greatly reduces the chance of cross-hybridization. Next step involves measuring the intensity of the fluorescence emitted from the slide. This is accomplished by placing the slides in a scanner, allowing the scanner to measure the intensity of fluorescence emitted after being excited by the laser. After which the final step involves the measurement of gene expression levels [83].

In two-dye microarray, fluorophore dye labeled both target gene samples are hybridized to the same array. Normally, Cy3 (fluorescence at red wavelength) and Cy5 (fluorescence at green wavelength) dyes are used to label the samples. Comparing the intensities of the fluorescence emitted by each wavelength in the array determines comparative expression levels between the samples of interest . Therefore this method offers reduced variability by performing direct

comparison between two samples [84]. In the same manner, one-dye platforms utilize fluorescence to label the samples but in this case only one dye is used. In this both the samples are placed in different arrays and expression level of each sample is determined by measuring the fluorescence [83]. One-dye systems present a very simplistic and flexible method to compare results between various groups of samples [80].

BeadChip microarray is designed by Illumina BeadArray Technology. The arrays designed comprise of thousands of 50mer oligonucleotide arranged in a unique bead type structure that are assembled into the microwells fixed onto surface of the BeadChip. Alongside expression measurement BeadChips can be used effectively for genotyping also. Different techniques can be utilized to synthesize probes of varying length: short (25-30 bases) and long (50-80 bases). An array contains approximately 30 instances of each bead type [85].

Affymetrix Gene And Exon Platform

Among all the microarray techniques available in the market, Affymetrix platform is the one most extensively accepted and utilized microarray technology. Manufactured by utilizing the process of photolithography and combinatorial chemistry, each GeneChip presents some 1.4 million individual oligonucleotide probes. The advantageous property of the spots on GeneChip is that each spot can contain millions of oligonucleotide copies [104]. Repeated illumination is performed to synthesize the probes onto the glass substrate. The glass substrate is layered with linkers, containing photoliable protecting groups. Normally, probe sets representing the genes or mRNAs of interest are 11-20 25mer oligonucleotide probe pairs. Based on the thier characteristics probe pairs are divided into two categories: perfect match (PM) probe and

mismatch (MM) probe (including modification on the 13th base) accounting for removal of non-specific hybridization and background noise [86].

Affymetrix gene array platform was designed to quantify the expression levels of well annotated genes obtained from any tissue. The probe set utilized for quantifying the expression levels of the gene consist of multiple probes that are complementary to different locations of the genomic locus. This platform contains 764,885 25-mer distinct probes that allow cross-examination of expression levels across 28,869 genes, based on March 2006 (UCSC hg18, NCBI Build 36). On average each gene contains 26 probes spanning different regions of the gene. This platform provides with an accurate and robust approach to detect transcriptional activities of genes [105].

GeneChip® Human Exon 1.0 ST Array is the latest platform developed by the Affymetrix for interrogating expression of genes at exon level [87]. The main objective behind developing exon array was "in interrogate each potential exon with one probe set over the entire genome on a single array". The exon array platform consists of more than 1.4 million probe sets, where on average each probesets consists of 4 probes, built using human genome assembly (July 2003, hg16, build 34) [106]. The probe sets corresponds to 1,796,124 probe selection region (PSR) from 1,084,639 exon clusters and more than 24,000 genes in human genome. Different annotation were used to support probe sets, 50% probe sets are based on single type annotation where half were derived from EST's and the other half from GENSCAN. The advantage of using exon array platform is that it allows detection of transcript diversity over a wide range. Using splice junction, detection of a small variation involving shift of 3 b.p. in splice sites becomes possible. However, this array platform as yet does not support exon junctions due to limited understanding of the variants present in the transcripts [106]. Increased probe concentration increases the density of probes four times and an eight-fold increase in perfect matches, as

compared to the previous platforms. Exon array utilizes specific probes instead of MM probes for detecting hybridization due to pure background. All these reformed advancements result in increased genomic coverage that provide with better estimation of gene-level expression analysis & also help to detect novel transcript variants due differences in exon-level expression [87].

Data Processing And Normalization

Processing of raw gene expression data obtained from any technique with an image analysis program is essential, as the initial data obtained is in form of scanned images. The processing of fluorescence emitted by probe in the array along with transformation of data need to be performed prior to applying statistical analyses on the data. The science behind normalization is to adjust for the variability encountered due to differences in microarray techniques and to reduce the background noise [88]. Signals emitting fluorescence are produced from the GeneChips during hybridization. The raw data from these signals is stored in a DAT extension file [89]. Information pertaining to image, pixels and technical information from the complete experiment is contained in the raw data. After performing grid alignment for registering set of unevenly spaced, parallel and perpendicular lines and computation of 75^{th} percentile of the spot specific pixel intensity, estimated intensities specific to each spot are stored in a CEL extension file. Since all the information present in DAT file is summarized to CEL file, grid alignment becomes an important step for avoiding errors and correctly summarizing information from DAT to CEL file. Additionally, image quality is affected by many other factors like flagging and background variations [89]. Another aspect that poses a significant affect over the quality of the image resulting in production of blurry images is the pixel in high intensity parts. These affect the reading of neighboring pixels possessing lower signals by recording the intensities from

different pixels. Non-specific hybridization due to MM probe also contributes to array's cell signal intensity. Therefore, to have more accurate estimate of cell-hybridization, background signals are subtracted. Similarly flagged features, bad features (higher pixel SD), negative feature (higher background compared to foreground) and dark feature (extremely low signal) all need to be removed manually or through computational means from the data [80].

Variations such as microarray manufacturing process, biological sample preparation and intensity measurements affect the data analysis. Therefore, normalization of the probe intensity data is performed which takes into account variations produced due to systemic errors and bias originating from microarrays and attempts at reducing the affect instilled by these variations [90]. Numerous statistical approaches to normalize the expression data have been proposed: Affymetrix Microarray Suite MAS5.0 software, Robust Multi-array Average (RMA) [91, 115] and GeneChip RMA (GCRMA) [92] are some widely used methods.

MAS5.0 utilizes intensity values of PM_{ij} and MM_{ij}, of the ith array and jth probe, to reduce the overall background noise. Implementation of log-transformation in the methods is responsible for reducing the reliance of variance on mean. The outliers are accommodated by the usage of Tukey's biweight function (T_{bi}) [91, 92]. Another method developed for normalization of data is RMA, which was mainly developed to overcome issues related to MM, as RMA considers information related to MM as biologically and statistically insignificant. Thus MM probes are not considered and PM_{ij} intensities are stated as $T(PM_{ij})$. The values are transformed by applying background corrections, normalization and applying logarithm [91]. To the background-corrected PM intensities base 2 logarithm is applied and substitution of original values is accomplished through utilization of mean quantile. This quantile normalization is performed at probe level, where the probe intensities for each array in array sets are distributed in similar

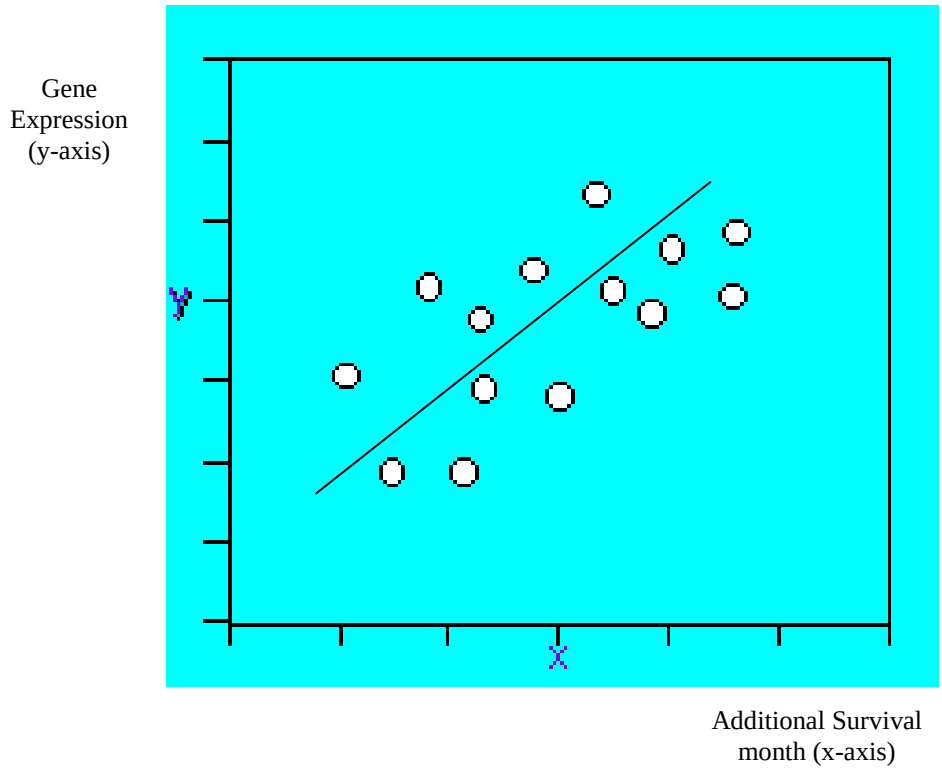
manner. Another normalization technique that is more extensively used by the scientific community is GCRMA. It is very much similar to RMA with the exception that it takes probe sequence into consideration, uses a different background correction and uses. Probe sequence consideration by GCRMA allows for intensity adjustment of probes showing different log intensities due to variation in GC content of the probes [92]. Thus PM values are corrected based on both GC content and MM probes, allowing GCRMA to have increased accuracy to estimate specific probe binding [92].

Comparison between the three techniques has established that GCRMA normalization method outshines the other two normalization techniques and RMA yield more promising results than MAS5.0 [92].

Identification Of Differential Expression Using Linear Models

General Linear Model (GLM) is one of the most commonly employed statistical models in the fields of social sciences and quantitative biology and comprises most of statistical analysis including the t-test, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), regression analysis, and many of the multivariate methods including factor analysis, cluster analysis, multidimensional scaling, discriminant function analysis, canonical correlation, and others. GLM describes the relationship between variables of interest by approximating to an optimal solution. The two variable case is the most simplistic type of GLM, consisting of explanatory and response variables.

Figure 1.3: Bivariate plot



Considering an example of tumor suppressor gene, looking at the plot it is expected that a positive relationship exists between the two variables with an increase in gene expression expected as the survival increases. How can we model this relationship so that we are able to describe the effect of survival accurately for any given sets of genes? The answer lies in the GLM approach which takes into account the effect of survival (X or explanatory variable) over gene expression (Y or response variable) and models this effect by fitting a line:

$$\mathbf{y = b_0 + bx + e}$$

y = a set of outcome variables
x = a set of pre-program variables or covariates
$\mathbf{b_0}$ = the set of intercepts (value of each y when each x=0)
b = a set of coefficients, one each for each x

We begin by collecting data for a number of individuals with varying survival periods. These individuals are called the experimental units. We plot the data into two-dimensional space and observe that both survival and gene expression are positively correlated. Now, we can draw a line that explains the 'general' effect of X over Y. By general, we mean that with a unit increase in X, we expect some units increase (or decrease) in Y. This quantity is called as the 'slope' of the fitted line. Because we are able to draw a 'linear' line (not a curve or any other shape), this generalization is referred to as the General Linear Model. The important point is that the effects of any sets of variables can be modeled quite precisely if we can fit a linear line that describes the relationship between the variables truthfully. It is not expected that all the points would fall directly onto the line. Usually, there will be scattering of data points around the fitted line (Figure 1.3). This realization is true because not every individual will have the same gene expression at any given time point. This scattering explains another important term in the GLM, which is the 'error' (e) term or the deviation of the experimental units around the mean. Ideally, we would like to minimize the experimental error and the line that best fits the data (with

minimum scattering) is the linear model for the given data. This additional error term in the GLM proves useful for explaining the variability in the data set and also tests whether a more complex model is needed or not (quadratic or cubic relationship). Model specification is an important step for correctly answering the research question and selection of insignificant variables can instill biasness in estimates of coefficients [103].

Functional Analysis

Gene expression patterns across thousands of genes results in production of tremendous amount of data which then needs to be analyzed to identify genes that are calculated to be significantly associated to the conditions under consideration. Various public databases, after being provided with gene list, categorize genes and gene products by including them into specific categories and sub-categories. These databases detect the genomic and functional categories enriched in the genelist provided, through help of statistical tools embedded in the database system [93]. Two such databases that are of high significance in biological community include Gene Ontology (GO) (http://www.geneontology.org/) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/pathway.html). Developed in 1998, the main objective of Gene ontology (GO) consortium was development of a system that provided a unified procedure to impart vocabulary to all novel findings even as knowledge of genes and proteins for all eukaryotic cells is accumulating and changing [94, 95]. Initially, GO contained information database against three model organism including mouse, yeast and fly. However, presently GO consortium constitutes major repositories microbial, plant and animal kingdoms genomes. GO web based tool contains diverse description on three categories including cellular component, molecular function and biological processes, where description in the sets is called ‘term’ [96].

The cellular category of GO consortium mainly deals with the exact location of gene product within a cell. The biochemical properties and activities related to genes are represented by molecular functions. A broad variety of functions both simple and complex are included in molecular function category. Lastly, biological objectives referring to processes including or leading to chemical and physical alterations of genes and their products are included in biological processes.

KEGG database is an online public resource containing information from 19 different databases on system, genomics and chemical properties across several genes and gene products. System information, one of three main categories, represents the functionality of any biological system. The second category, "Genomic information" represents the genomic building blocks of life while Chemical information represents the chemical building blocks necessary for life [97]. The graphic tool embedded in the KEGG database system allows retrieval of information pertaining to molecular networks and cellular processes. The graphic tool illustrates many metabolic interaction networks, genetic and environmental information processing, chemical structure transformation network and human diseases. The pathway maps, representing a particular network are described through nodes (genes, proteins) and edges (relation, reaction).

Computational and statistical analysis conducted on large scale genomic data results in formation of genelist that includes genes found to be significantly associated with the characteristic of interest. Making biological inferences from the significant genes requires functional categorization of these genes in groups based on some specific pattern. Over the years some 68 reputable bioinformatics tools have been developed that take into consideration biological information accumulated in various public databases (GO, KEGG). Utilizing these tools allows systematic assembly of enriched functions and pathways corresponding to the significant genes

[98]. Some of these high-throughput enrichment tools include Onto-Express, MAPPFinder, GoMiner, DAVID, EASE, GeneMerge and FuncAssociate that extract relevant GO terms and KEGG pathways from several databases. One of these online resources Database for Annotation, Visualization and Integrated Discovery (http://david.abcc.ncifcrf.gov/) (DAVID) is a highly reputable functional enrichment tool, which amalgamates/accumulates various features including back-end annotation database, advanced enrichment algorithm and powerful exploratory data mining ability within itself. DAVID utilizes information from several databases to extract GO terms and KEGG pathways, Fischer's Exact test is then performed to categorize the genelists [99]. Thus enrichment analysis provides valuable information related to gene ontologies and pathways for the genelist provided.

Transcriptome variants produced due to alternative exon usage (AEU) are an important aspect of gene regulation. Almost 90 % of the multi-exon genes in humans are transcribed into multiple transcript variants as result of alternative mRNA splicing [100]. Alternative exon usage has been associated with proliferation of malignant cells in humans [57, 101]. Glioblastoma multiforme (GBM) is the most severe form of malignant brain tumors and the expression of numerous genes has been associated with this cancer [62]. These transcripts might be related to a specific metastatic phenotype and can potentially function as diagnostic and prognostic biomarkers and even as therapeutic drug targets [102]. However, few consistent gene expression biomarkers of GBM have been reported [57]. Two reasons for this are the limited consideration of AEU and, the analytical approaches typically used to study AEU that ignore the relationship between exons within a gene. The goals of this study are to develop a general hierarchical model to identify the differential associations between cancer survival and expression at a gene or exon level that indicate AEU and to apply this methodology to identify biomarkers of GBM survival.

References

1. White RJ, Sharrocks AD: **Coordinated control of the gene expression machinery.** Trends Genet 2010, **26**(5):214-220.

2. Crick F: **Central dogma of molecular biology.** Nature 1970, **227**(5258):561-563.

3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al: **The sequence of the human genome.** Science 2001, **291**(5507):1304-1351.

4. Laderas TG, Walter NA, Mooney M, Vartanian K, Darakjian P, Buck K, Harrington CA, Belknap J, Hitzemann R, McWeeney SK: **Computational detection of alternative exon usage.** Front Neurosci 2011, **5**:69.

5. Modrek B, Lee C: **A genomic view of alternative splicing.** Nat Genet 2002, **30**(1):13-19.

6. Matlin AJ, Clark F, Smith CW: **Understanding alternative splicing: towards a cellular code.** Nat Rev Mol Cell Biol 2005, **6**(5):386-398.

7. Sanford JR, Caceres JF: **Pre-mRNA splicing: life at the centre of the central dogma.** J Cell Sci 2004, **117**(Pt 26):6261-6263.

8. Barash Y, Blencowe BJ, Frey BJ: **Model-based detection of alternative splicing signals.** Bioinformatics 2010, **26**(12):i325-33.

9. Black DL: **Mechanisms of alternative pre-messenger RNA splicing.** Annu Rev Biochem 2003, **72**:291-336.

10. Sharma S, Maris C, Allain FH, Black DL: **U1 snRNA directly interacts with polypyrimidine tract-binding protein during splicing repression.** Mol Cell 2011, **41**(5):579-588.

11. Zhou HL, Lou H: **Repression of prespliceosome complex formation at two distinct steps by Fox-1/Fox-2 proteins.** Mol Cell Biol 2008, **28**(17):5507-5516.

12. Spellman R, Smith CW: **Novel modes of splicing repression by PTB.** Trends Biochem Sci 2006, **31**(2):73-76.

13. Nasim FU, Hutchison S, Cordeau M, Chabot B: **High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism.** RNA 2002, **8**(8):1078-1089.

14. Martinez-Contreras R, Fisette JF, Nasim FU, Madden R, Cordeau M, Chabot B: **Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing.** PLoS Biol 2006, **4**(2):e21.

15. Chen M, Manley JL: **Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.** Nat Rev Mol Cell Biol 2009, **10**(11):741-754.

16. Yu Y, Maroney PA, Denker JA, Zhang XH, Dybkov O, Luhrmann R, Jankowsky E, Chasin LA, Nilsen TW: **Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition.** Cell 2008, **135**(7):1224-1236.

17. 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.

18. Zhu J, Krainer AR: **Pre-mRNA splicing in the absence of an SR protein RS domain.** Genes Dev 2000, **14**(24):3166-3178.

19. Crawford JB, Patton JG: **Activation of alpha-tropomyosin exon 2 is regulated by the SR protein 9G8 and heterogeneous nuclear ribonucleoproteins H and F.** Mol Cell Biol 2006, **26**(23):8791-8802.

20. Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH: **An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells.** Nat Struct Mol Biol 2009, **16**(2):130-137.

21. Mauger DM, Lin C, Garcia-Blanco MA: **hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 exon IIIc.** Mol Cell Biol 2008, **28**(17):5403-5419.

22. Dredge BK, Stefani G, Engelhard CC, Darnell RB: **Nova autoregulation reveals dual functions in neuronal splicing.** EMBO J 2005, **24**(8):1608-1620.

23. Dredge BK, Darnell RB: **Nova regulates GABA(A) receptor gamma2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer.** Mol Cell Biol 2003, **23**(13):4687-4700.

24. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB: **An RNA map predicting Nova-dependent splicing regulation.** Nature 2006, **444**(7119):580-586.

25. Venables JP, Tazi J, Juge F: **Regulated functional alternative splicing in Drosophila.** Nucleic Acids Res 2012, **40**(1):1-10.

26. Henkin TM: **Riboswitch RNAs: using RNA to sense cellular metabolism.** Genes Dev 2008, **22**(24):3383-3390.

27. Cheah MT, Wachter A, Sudarsan N, Breaker RR: **Control of alternative RNA splicing and gene expression by eukaryotic riboswitches.** Nature 2007, **447**(7143):497-500.

28. Kishore S, Stamm S: **The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C.** Science 2006, **311**(5758):230-232.

29. Su WL, Modrek B, GuhaThakurta D, Edwards S, Shah JK, Kulkarni AV, Russell A, Schadt EE, Johnson JM, Castle JC: **Exon and junction microarrays detect widespread mouse strain- and sex-bias expression differences.** BMC Genomics 2008, **9**:273.

30. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP: **FIRMA: a method for detection of alternative splicing from exon array data.** Bioinformatics 2008, **24**(15):1707-1714.

31. Cline MS, Blume J, Cawley S, Clark TA, Hu JS, Lu G, Salomonis N, Wang H, Williams A: **ANOSVA: a statistical method for detecting splice variation from expression data.** Bioinformatics 2005, **21 Suppl 1**:i107-15.

32. Shai O, Morris QD, Blencowe BJ, Frey BJ: **Inferring global levels of alternative splicing isoforms using a generative model of microarray data.** Bioinformatics 2006, **22**(5):606-613.

33. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** Proc Natl Acad Sci U S A 1998, **95**(25):14863-14868.

34. Fagnani M, Barash Y, Ip JY, Misquitta C, Pan Q, Saltzman AL, Shai O, Lee L, Rozenhek A, Mohammad N, Willaime-Morawek S, Babak T, Zhang W, Hughes TR, van der Kooy D, Frey BJ, Blencowe BJ: **Functional coordination of alternative splicing in the mammalian central nervous system.** Genome Biol 2007, **8**(6):R108.

35. Zheng H, Hang X, Zhu J, Qian M, Qu W, Zhang C, Deng M: **REMAS: a new regression model to identify alternative splicing events from exon array data.** BMC Bioinformatics 2009, **10 Suppl 1**:S18.

36. Dralyuk I, Brudno M, Gelfand MS, Zorn M, Dubchak I: **ASDB: database of alternatively spliced genes.** Nucleic Acids Res 2000, **28**(1):296-297.

37. Ji H, Zhou Q, Wen F, Xia H, Lu X, Li Y: **AsMamDB: an alternative splice database of mammals.** Nucleic Acids Res 2001, **29**(1):260-263.

38. Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick JS: **ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome.** Nat Genet 2000, **24**(4):340-341.

39. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** Nucleic Acids Res 2001, **29**(13):2850-2859.

40. Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA: **ASD: a bioinformatics resource on alternative splicing.** Nucleic Acids Res 2006, **34**(Database issue):D46-55.

41. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation.** Genome Biol 2006, **7 Suppl 1**:S12.1-14.

42. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing.** Gene 2005, **344**:1-20.

43. Tazi J, Bakkour N, Stamm S: **Alternative splicing and disease.** Biochim Biophys Acta 2009, **1792**(1):14-26.

44. Camacho-Vanegas O, Narla G, Teixeira MS, DiFeo A, Misra A, Singh G, Chan AM, Friedman SL, Feuerstein BG, Martignetti JA: **Functional inactivation of the KLF6 tumor suppressor gene by loss of heterozygosity and increased alternative splicing in glioblastoma.** Int J Cancer 2007, **121**(6):1390-1395.

45. Kim E, Goren A, Ast G: **Insights into the connection between cancer and alternative splicing.** Trends Genet 2008, **24**(1):7-10.

46. Grabowski PJ, Black DL: **Alternative RNA splicing in the nervous system.** Prog Neurobiol 2001, **65**(3):289-308.

47. Ehlers MD, Tingley WG, Huganir RL: **Regulated subcellular distribution of the NR1 subunit of the NMDA receptor.** Science 1995, **269**(5231):1734-1737.

48. Ehlers MD, Fung ET, O'Brien RJ, Huganir RL: **Splice variant-specific interaction of the NMDA receptor subunit NR1 with neuronal intermediate filaments.** J Neurosci 1998, **18**(2):720-730.

49. Wu W, Sato K, Koike A, Nishikawa H, Koizumi H, Venkitaraman AR, Ohta T: **HERC2 is an E3 ligase that targets BRCA1 for degradation.** Cancer Res 2010, **70**(15):6384-6392.

50. Schwerk C, Schulze-Osthoff K: **Regulation of apoptosis by alternative pre-mRNA splicing.** Mol Cell 2005, **19**(1):1-13.

51. Ayroldi E, D'Adamio F, Zollo O, Agostini M, Moraca R, Cannarile L, Migliorati G, Delfino DV, Riccardi C: **Cloning and expression of a short Fas ligand: A new alternatively spliced product of the mouse Fas ligand gene.** Blood 1999, **94**(10):3456-3467.

52. Pearn J: **Classification of spinal muscular atrophies.** Lancet 1980, **1**(8174):919-922.

53. Zhang Z, Lotti F, Dittmar K, Younis I, Wan L, Kasim M, Dreyfuss G: **SMN deficiency causes tissue-specific perturbations in the repertoire of snRNAs and widespread defects in splicing.** Cell 2008, **133**(4):585-600.

54. Ajmal M, Ahmed W, Akhtar N, Sadeque A, Khalid A, Benish Ali SH, Ahmed N, Azam M, Qamar R: **A novel pathogenic nonsense triple-nucleotide mutation in the low-density lipoprotein receptor gene and its clinical correlation with familial hypercholesterolemia.** Genet Test Mol Biomarkers 2011, **15**(9):601-606.

55. Zhu H, Tucker HM, Grear KE, Simpson JF, Manning AK, Cupples LA, Estus S: **A common polymorphism decreases low-density lipoprotein receptor exon 12 splicing efficiency and associates with increased cholesterol.** Hum Mol Genet 2007, **16**(14):1765-1772.

56. Zou F, Gopalraj RK, Lok J, Zhu H, Ling IF, Simpson JF, Tucker HM, Kelly JF, Younkin SG, Dickson DW, Petersen RC, Graff-Radford NR, Bennett DA, Crook JE, Younkin SG, Estus S: **Sex-dependent association of a common low-density lipoprotein receptor polymorphism with RNA splicing efficiency in the brain and Alzheimer's disease.** Hum Mol Genet 2008, **17**(7):929-935.

57. 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.

58. Venables JP: **Aberrant and alternative splicing in cancer.** Cancer Res 2004, **64**(21):7647-7654.

59. Fackenthal JD, Godley LA: **Aberrant RNA splicing and its functional consequences in cancer cells.** Dis Model Mech 2008, **1**(1):37-42.

60. Serao NV, Delfino KR, Southey BR, Beever JE, Rodriguez-Zas SL: **Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival.** BMC Med Genomics 2011, **4**:49.

61. Holland EC: **Glioblastoma multiforme: the terminator.** Proc Natl Acad Sci U S A 2000, **97**(12):6242-6244.

62. Ohgaki H, Dessen P, Jourde B, Horstmann S, Nishikawa T, Di Patre PL, Burkhard C, Schuler D, Probst-Hensch NM, Maiorka PC, Baeza N, Pisani P, Yonekawa Y, Yasargil MG, Lutolf UM, Kleihues P: **Genetic pathways to glioblastoma: a population-based study.** Cancer Res 2004, **64**(19):6892-6899.

63. Ohgaki H, Kleihues P: **Genetic pathways to primary and secondary glioblastoma.** Am J Pathol 2007, **170**(5):1445-1453.

64. Johnson DR, O'Neill BP: **Glioblastoma survival in the United States before and during the temozolomide era.** J Neurooncol 2012, **107**(2):359-364.

65. Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** Nature 2008, **455**(7216):1061-1068.

66. Chen P, Aldape K, Wiencke JK, Kelsey KT, Miike R, Davis RL, Liu J, Kesler-Diaz A, Takahashi M, Wrensch M: **Ethnicity delineates different genetic pathways in malignant glioma.** Cancer Res 2001, **61**(10):3949-3954.

67. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, Kleihues P: **The 2007 WHO classification of tumours of the central nervous system.** Acta Neuropathol 2007, **114**(2):97-109.

68. Sjostrom S, Andersson U, Liu Y, Brannstrom T, Broholm H, Johansen C, Collatz-Laier H, Henriksson R, Bondy M, Melin B: **Genetic variations in EGF and EGFR and glioblastoma outcome.** Neuro Oncol 2010, **12**(8):815-821.

69. Bhowmick DA, Zhuang Z, Wait SD, Weil RJ: **A functional polymorphism in the EGF gene is found with increased frequency in glioblastoma multiforme patients and is associated with more aggressive disease.** Cancer Res 2004, **64**(4):1220-1223.

70. Ohgaki H, Kleihues P: **Genetic alterations and signaling pathways in the evolution of gliomas.** Cancer Sci 2009, **100**(12):2235-2241.

71. Zhao S, Lin Y, Xu W, Jiang W, Zha Z, Wang P, Yu W, Li Z, Gong L, Peng Y, Ding J, Lei Q, Guan KL, Xiong Y: **Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1alpha.** Science 2009, **324**(5924):261-265.

72. Ramskold D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.** PLoS Comput Biol 2009, **5**(12):e1000598.

73. Lo HW, Zhu H, Cao X, Aldrich A, Ali-Osman F: **A novel splice variant of GLI1 that promotes glioblastoma cell migration and invasion.** Cancer Res 2009, **69**(17):6790-6798.

74. Izaguirre DI, Zhu W, Hai T, Cheung HC, Krahe R, Cote GJ: **PTBP1-dependent regulation of USP5 alternative RNA splicing plays a role in glioblastoma tumorigenesis.** Mol Carcinog 2011, .

75. Montgomery SB, Dermitzakis ET: **The resolution of the genetics of gene expression.** Hum Mol Genet 2009, **18**(R2):R211-5.

76. Larson DR, Singer RH, Zenklusen D: **A single molecule view of gene expression.** Trends Cell Biol 2009, **19**(11):630-637.

77. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** Science 1995, **270**(5235):467-470.

78. Alwine JC, Kemp DJ, Stark GR: **Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.** Proc Natl Acad Sci U S A 1977, **74**(12):5350-5354.

79. Becker-Andre M, Hahlbrock K: **Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY).** Nucleic Acids Res 1989, **17**(22):9437-9446.

80. Petersen D, Chandramouli GV, Geoghegan J, Hilburn J, Paarlberg J, Kim CH, Munroe D, Gangi L, Han J, Puri R, Staudt L, Weinstein J, Barrett JC, Green J, Kawasaki ES: **Three microarray platforms: an analysis of their concordance in profiling gene expression.** BMC Genomics 2005, **6**:63.

81. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** Nat Rev Genet 2009, **10**(1):57-63.

82. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A et al: **Standardizing global gene expression analysis between laboratories and across platforms.** Nat Methods 2005, **2**(5):351-356.

83. Stears RL, Martinsky T, Schena M: **Trends in microarray analysis.** Nat Med 2003, **9**(1):140-145.

84. Tang T, Francois N, Glatigny A, Agier N, Mucchielli MH, Aggerbeck L, Delacroix H: **Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment.** Bioinformatics 2007, **23**(20):2686-2691.

85. Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S, Bennett H, Rigault P, Barker D, McDaniel TK, Chee MS: **A novel, high-performance random array platform for quantitative gene expression profiling.** Genome Res 2004, **14**(11):2347-2356.

86. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** Nucleic Acids Res 2003, **31**(4):e15.

87. Kapur K, Xing Y, Ouyang Z, Wong WH: **Exon arrays provide accurate assessments of gene expression.** Genome Biol 2007, **8**(5):R82.

88. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** Nucleic Acids Res 2002, **30**(4):e15.

89. Arteaga-Salas JM, Zuzan H, Langdon WB, Upton GJ, Harrison AP: **An overview of image-processing methods for Affymetrix GeneChips.** Brief Bioinform 2008, **9**(1):25-33.

90. Do JH, Choi DK: **Normalization of microarray data: single-labeled and dual-labeled arrays.** Mol Cells 2006, **22**(3):254-261.

91. 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.

92. Vardhanabhuti S, Blakemore SJ, Clark SM, Ghosh S, Stephens RJ, Rajagopalan D: **A comparison of statistical tests for detecting differential expression using Affymetrix oligonucleotide microarrays.** OMICS 2006, **10**(4):555-566.

93. Dopazo J: **Functional interpretation of microarray experiments.** OMICS 2006, **10**(3):398-410.

94. Reference Genome Group of the Gene Ontology Consortium: **The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species.** PLoS Comput Biol 2009, **5**(7):e1000431.

95. Gene Ontology Consortium: **The Gene Ontology: enhancements for 2011.** Nucleic Acids Res 2012, **40**(Database issue):D559-64.

96. Pal D: **On gene ontology and function annotation.** Bioinformation 2006, **1**(3):97-98.

97. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** Nucleic Acids Res 2010, **38**(Database issue):D355-60.

98. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward comprehensive functional analysis of large gene lists.** Nucleic Acids Res 2009, **37**(1):1-13.

99. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** Nat Protoc 2009, **4**(1):44-57.

100. Kim N, Lee C: **Bioinformatics detection of alternative splicing.** Methods Mol Biol 2008, **452**:179-197.

101. Lin E, Li L, Guan Y, Soriano R, Rivers CS, Mohan S, Pandita A, Tang J, Modrusan Z: **Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers.** Mol Cancer Res 2009, **7**(9):1466-1476.

102. Robinson MD, Speed TP: **Differential splicing using whole-transcript microarrays.** BMC Bioinformatics 2009, **10**:156.

103. Wojtek, J. and Krzanowski: **An introduction to Statistical Modeling**. Arnold texts in statistics, ed. D. Collett. 1998, London: Arnol.

104. Affymetrix: **GeneChip® expression analysis technical manual**. 2005.

105. Affymetrix: **GeneChip® Gene 1.0 ST Array Design**. 2007.

106. Affymetrix: **GeneChip® Exon 1.0 Array Design**. 2005.

CHAPTER 2

RESEARCH PAPER

Background

Alternative splicing (AS) is characterized by the formation of different mRNA isoforms as a result of including or excluding different exonic or intronic segments. This process is responsible for generating protein diversity [1-3]. AS can be divided into three broad categories; intron retention, cryptic splice site usage (functions by elongating or shortening the exon), and alternative exon usage (AEU) or exon skipping. AEU includes cassette exons (discrete exons that can be independently included or excluded) and mutually exclusive splicing (which involves the selection of one from a group of exon variants) [3]. Approximately 75% of multi-exon genes exhibit AS in humans [4] and on average more than 3 alternative transcripts are mapped to a gene [5]. The identification of "exon-level" expression profiles and characterization of AS events has become possible with the availability of exon platforms (e.g. GeneChip Exon Array).

The brain exhibits particularly high rates of AS [6] and the highest number of AEU events [7]. Regulation of gene expression due to splicing has been associated with cancer. Many AEU events have been associated with disordered cell differentiation and signaling that contribute to stem cell like proliferation of cancer cells [8].

Glioblastoma multiforme (GBM) is an aggressive type of brain cancer and the role of genes and AEU on GBM survival is still not completely understood [9-11]. Most work on AS and GBM studied individual genes or compared AS between GBM and control (e.g. blood) samples. The relationship between AS and the survival of individuals diagnosed with GBM has not been studied. Understanding of the factors influencing survival is particularly important in GBM

because the average survival after diagnostic is approximately one year [12, 13]. Furthermore, several epidemiological factors influence GBM survival including gender, race and treatment [14]. Thus, a more accurate understanding of the relationship between AS and GBM survival must consider epidemiological factors and inter-individual variability.

Several approaches to identify AS events have been proposed. However, most approaches have limitations that can bias the identification and characterization of AEU. For example, Su et al. developed an individual exon approach [15] that does not model the covariation between exons within a gene. Purdom et al. used the residuals from probe level analysis to identify AEU on a per-sample level [16]. The sample-level analysis challenges the detection of AEU events or the identification of common patterns across patients receiving the same treatment or from the same epidemiological strata. Laderas et al. and Zheng et al. proposed group comparison using linear models to overcome the limitations of the previous approach [3, 17]. However, group comparison is not suited to identify AEU associated with other conditions such as survival or time-to-event. In addition, the previous implementation does not account for correlations between exons measured on the same sample. Cline et al. formulated a ANalysis Of Splice VAriation ANOSVA approach that cannot be used in gene that produced more than one splice form [18].

The goal of this study is to implement an exon-based and gene-centric model that allows the detection of AEU associated with cancer survival. The approach addresses the limitations of previous approaches by modeling the exon-level expression profiles within gene from all samples across all treatments or conditions studied. Our approach accommodates the dependencies between exons within a gene and patient and allows testing the hypothesis of differential exon expression or usage between treatment groups. A unique advantage of our

flexible approach is that one model encompasses all scenarios: i) multi-exon genes that have AEU, ii) multi-exon genes that do not have AEU, and iii) single-exon genes. This novel approach was applied to the identification of AEU associated with GBM survival. The performance of the approach and accuracy of the results were assessed by using separate training and validation data sets. Gene set enrichment and gene functional analyses offered insights into the biological processes related to the AEU genes associated with survival. Results were mapped to genes and to known or predicted AS events to further confirm and add biological interpretation to the results of our model.

Materials and Methods

Training Dataset

Survival, clinical and exon expression information from 250 patients diagnosed with GBM was obtained from The Cancer Genome Atlas (http://cancergenome.nih.gov/) repository (May 2011 data freeze). Surgical samples had a minimum of 80% tumor nuclei and maximum of 50% necrosis. In this training data set, clinical or epidemiological variables considered in the analysis of exon expression included treatment (levels: chemo-radiation-targeted or CRT, chemo-radiation-non targeted or CRnT, radiation or R, other therapies or OTHER or no therapy or NONE); racial ethnicity (white Caucasian and others); and gender (male and female). These clinical factors were accounted for in the model because of their known association with survival [9]. The survival variable associated with exon expression was the time from diagnosis to death (expressed in months).

Exon expression measurements from each patient were obtained using the AffymetrixGeneChip® Human Exon 1.0 ST Array. Platform details can be found at (http://www.affymetrix.com/support/technical/annotationfilesmain.affx) (Affymetrix, Inc., 2012). Briefly, this platform includes information from 1432143 (1.4 million) probe sets representing known and predicted exons on both strands of the genome. These probe sets were mapped to more than 25000 genes. Array platform and data was log-2 transformed and normalized using quantile and RMA normalization at the probe level following the procedures described in Beehive (http://stagbeetle.animal.uiuc.edu/Beehive). Probes sets within exon were collapsed using a Tukey biweight function that provides an iterative reweighed measure of central tendency. This robust statistic provides a single exon expression that is not heavily influenced by extreme probe expression levels [19].

Model

One general exon expression model was developed to describe the association between and exon expression and GBM survival adjusted for other clinical factors. Three specifications of the model accommodated three groups of genes: 1) multi-exon genes exhibiting AEU, 2) multi-exon genes with no evidence of AEU, and 3) single-exon genes. The general model is:

$$y_{ijklmn} = \mu + G_i + R_j + T_k + S_l + E_m + (SE)_{lm} + P_n + e_{ijklmn} \qquad [1]$$

where y_{ijklmn} denotes the expression of the mth exon (E_m), recorded on the nth patient (P_n) that has the ith gender (G_i), jth race\ethnicity (R_j), received the kth therapy (T_k), and survived lth months after diagnostic (S_l). In addition, e_{ijklmn} is the residual associated with the y_{ijklmn} observation and SE denotes the interaction between survival and exon on expression levels.

Fixed effects are G, R, T, and S. Random effects E, SE, and P are assumed each to be independent and follow a Gaussian distribution with its own variance.

Model [1] allows the study of two scenarios (groups 1 and 2). First, a significant SE effect constitutes evidence of an AEU scenario and thus differential survival across exons. This model can be used to identify AS biomarkers of GBM survival that exhibit AEU. Second, a significant S effect together with a non-significant SE effect constitutes evidence of a general association between gene expression and survival, regardless of exon. In addition, a significant E effect indicates that the exons have differential expression. However, the association between the exons expression and survival is similar because SE is non-significant. This result can be used to identify multi-exon biomarkers of GBM survival that do not exhibit AEU.

The specification for the single-exon gene (group 3) is a reduced version of the full multi-exon model [1] that excludes E and SE. Like with the multi-exon model under non-significant SE, a significant S effect is evidence of association between the single-exon gene expression and survival and can be used to identify single-exon biomarkers of GBM survival.

The novel gene-centered analysis allows accounting for the covariance between exon expression within a gene and the hierarchical nature of the model allows the inclusion of the covariance between exon expression within a patient. The analysis of expression data at the exon level permitted the identification of AEU by testing the null hypothesis of no differential association between expression and survival across exons within a gene.

False Discovery Rate adjustment (FDR) of the P-values allowed controlling for multiple testing (Benjamini and Hochberg) [20]. In addition, a more stringent P-value threshold was considered for the detection of AEU associations with survival (significant SE) relative to the detection of a general association between gene expression and survival (significant S) in the multi-exon

scenarios. The more stringent P-values required for detection of AEU accounted for the multiple comparisons of the survival-expression associations among potentially numerous exons. A separate FDR adjustment of the P-values from the single-exon analysis was implemented because of the different number of parameters between the multi- and single-exon models. The P-value < 5.0E-4 corresponds to a FDR adjusted P-value < 5.0E-2or multi exon genes and to a FDR adjusted P-value < 1.0E-1for genes with single exon.

Functional and pathway analyses of the genes exhibiting significant evidence (P-value < 5.0E-4) of AEU associated with GBM survival used hypergeometric tests and was implemented in DAVID (http://david.abcc.ncifcrf.gov/) [21, 22]. Gene set enrichment analysis of the association between expression and GBM survival among all the genes studied in the platform followed the approach described by Subramanian et al. [23] implemented in BABELOMICS 4.3 (http://babelomics.bioinfo.cipf.es/index.html, [24]. For this analysis, the association between each gene and survival was characterized by the estimate of change in expression per additional survival month standardized by the standard error of the estimate. The enrichment of Gene Ontology (GO; http://www.geneontology.org/) biological processes, molecular functions, and KEGG (http://www.genome.jp/kegg/pathway.html) pathways was investigated. Finally, P-values of the enriched categories were adjusted for multiple testing using the FDR correction.

We recognize that statistical evidence is one component in the identification of AEU. However, it is biologically unlikely that AS events skip single or two exons across a gene. In addition, changes in the association between exon expression and survival may be statistically significant due to the substantial number of patients studied but may only represent small fold changes.

Thus, three types of evidence were used to identify AEU. We looked for a) significant variations in the associations between exon expression and survival across a gene, b) consistent (over or

under-expressed) differential expression in multiple consecutive exons, and c) a minimum exon differential expression (< 0.995 or > 1.005 fold change / additional survival month). Consistent patterns of expression across consecutive exons were identified using a moving average analysis [25]. A moving average analysis that computes the average expression across multiple exons at a time was used to predict a continuous trajectory of exon expression across the gene. This moving average trend of exon expression across the gene facilitated the identification of consistent changes in the pattern of over or under-expression across the exons within a gene.

The exon expression estimates and the moving average trajectory of the estimates across individual genes were aligned to known or predicted alternative transcript variants reported in the AceView database (http://genome.ucsc.edu/cgibin/hgTrackUi?hgsid=243882995&g=acembly&hgTracksConfigPage=configure) that are available through the UCSC Genome Browser (http://genome.ucsc.edu). This visualization strategy facilitated the interpretation of results and the AS models offered an independent *in silico* confirmation of the AEU events identified.

Validation Dataset

Genes exhibiting AEU in the training data set were confirmed in an independent set of 78 patients obtained from the same repository. The reliability of the exon expression profiles associated with survival identified in the training data set was assessed using a two-stage approach. First, the parameter estimates (i.e. changes in exon expression per one month increase in survival) obtained from the analysis of the training data set were applied to the covariate information from the patients in the validation data set and predictions of exon expression were obtained. Second, the predicted exon expression values were compared to the corresponding

observed expression values. The performance of the training estimates was evaluated using the coefficient of determination (R^2) indicators [26]. High R^2 on the validation data set based on the training data set estimates indicate the reliability of the exon expression patterns identified.

Results and Discussion

Expression measurements of 269951 exons from 25403 genes were analyzed. Of these, 2857, 20288, 1965 and 293 genes had one, 2 to 24, 25 to 49 and 50 or more exons, respectively. The number of exons per gene ranged from 1 to 191 and the average number of exons was 10.75. Table 1 summarizes the distribution of the 250 and 78 individuals diagnosed with GBM analyzed in the training and validation data sets respectively, across clinical factors and survival descriptive statistics. The distribution of observations across clinical factors was consistent across data sets.

The results from training data set are summarized in three groups according to the model used and evidence supporting AEU: 1) multi-exon genes exhibiting exon dependent association with GBM survival or AEU (group 1 genes), 2) multi-exon genes exhibiting exon-independent association with GBM survival or no AEU (group 2 genes), and 3) single-exon genes exhibiting association with GBM survival (group 3 genes). The general model proposed supports the consistent analysis of single-exon and multi-exon genes and identifies gene or exon associations with survival. The general and hierarchical nature of the model permits testing a myriad of hypothesis. The consideration of an interaction between survival and exon allowed the identification of associations between particular exons and survival and corresponding characterization of AEU (exon-specific fold change /additional survival month).

Table 1. Distribution of patients with glioblastoma multiforme analyzed by level of clinical factors.

		Training data set		Validation data set	
		Number	Percentage	Number	Percentage
Patients		250	76.22	78	23.78
Race[1]	Caucasian	222	88.80	71	91.03
	Other	28	11.20	07	8.97
Gender	Females	94	37.60	29	37.18
	Males	156	62.40	49	62.82
Therapy[2]	R	63	25.20	21	26.92
	CRT	27	10.80	07	8.97
	CRnT	99	39.60	31	39.74
	OTHER	35	14.00	10	12.82
	NONE	26	10.40	09	11.54
Survival (months)		17.46	0.16 - 128	15.02	0.10 – 77.57

[1]Race: White Caucasian, and all other race-ethnicity groups

[2]Therapy; R: radiation therapy alone; CRnT: chemotherapy plus radiation and no targeted therapy; CRT: chemotherapy plus radiation and targeted therapy; OTHER: all other therapies None: no therapy.

In the event of non-significant exon-by-survival interaction, the inclusion of a general survival covariate allowed the detection of a general association between gene expression and survival that does not differ among exons. The novel modeling of exons as random effect levels permitted the specification of a variance-covariance structure between the exons within a gene. The removal of exon-dependent terms from the full multi-exon model offered a model suitable for single-exon genes. In addition, the block patient effect accommodates the covariance between exon levels measured in the same patients.

Multi-Exon Genes Exhibiting Exon-Dependent Association with Glioblastoma Multiforme Survival

At unadjusted P-value < 5.0E-4 (equivalent to FDR-adjusted P-value < 5.0E-2), 2477 multi-exon genes exhibited AEU associated with survival (group 1 genes), 24 multi-exon genes exhibited expression associated with survival albeit no evidence of AEU (group 2 genes), and 8 single-exon genes exhibited expression associated with survival (group 3 genes). At P-value < 1.0E-5, P-value < 1.0E-6, P-value < 1.0E-7, P-value < 1.0E-8, the number of genes exhibiting AEU (group 1 genes) were 592, 313, 201, and 129 respectively.

Table 2 summarizes the top 36 multi-exon genes that have the most significant (P-value < 1.0E-11) AEU or exon-dependent association with GBM survival (group 1 genes) due to space constraints.

The nature of the association between expression and GBM is characterized by the sign and value of the expression fold change per additional month in survival. Tables 2, 4, and 5 include a general gene-wise estimate of expression fold change per additional month in survival for multi-exon with (group 1) and without AEU (group 2) and single-exon (group 3) genes. The meaning

of this fold change estimate is straightforward for group 2 and 3 genes because these genes exhibit a single general association with GBM survival, Attention should be exercised when interpreting the general fold change estimate for group 1 AEU genes because these genes exhibit significant variation in association between expression and survival among exons.

The top 36 genes exhibiting significant evidence of AEU have a minimum of 90 exons (Table 2). This result suggests that genes with high number of exons are more likely to experience AEU events that influence GBM survival. It is unlikely that high number of exons biased the identification of AEU because a stringent P-value threshold was used.

Among the 36 genes that have significant AEU association with GBM survival most have been related to cancer. Of these, 10 genes including titin (*Ttn*), polycystic kidney disease 1 (*Pkd1*), spectrin repeat containing, nuclear envelope 1 (*Syne1*), small nuclear ribinucleoprotein (*Snrpn*), phosphodiesterase 4D interacting protein (*Pde4dip*), obscurin (*Obscn*), dystonin (*Dst*), microtubule-actin cross-linking factor 1 (*Macf1*), ryanodine receptor 1 (*Ryr1*) and ryanodine receptor 2 (*Ryr2*) have been previously associated with GBM. Additionally, 13 genes Smg-1 homolog (*Smg1*), Nebulin (*Neb*), TBC1 domain family, member 3 (*Tbc1d3*), Anaphase promoting complex subunit 1 (*Anapc1*), Spectrin repeat containing, nuclear envelope 1 (*Syne2*), Neuroblastoma breakpoint family, member 10 (*Nbpf10*), Mucin 19 (*Muc19*), Collagen, type VII, alpha 1 (*Col7a1*), Ubiquitin protein ligase E3 component n-recognin 4 (*Ubr4*), Hemicentin 1 (*Hmcn1*), Collagen, type IV, alpha 5 (*Col4a5*), Ryanodine receptor 3 (*Ryr3*), G protein-coupled receptor 98 (*Gpr98*) have been previously associated to cancers other than GBM. The list of references is summarized in Table 2. Additionally, literature review also supported the presence of AS in most of the genes.

Table 2. Top 36 multi-exon genes that have significant alternative exon usage associated with glioblastoma multiforme survival.

Gene Symbol	Estimate[1]	SE[2]	P-value AEU[3]	Fold change[4]	Exon Count[5]	Literature[6]
Ttn	0.0007	0.0001	4.2E-38	0.9993	340	[27] [G]
Smg1	0.0017	0.0002	2.0E-24	1.0001	209	[70] [AS]
Neb	0.0007	0.0001	3.2E-21	0.9973	180	[71, 72][C, AS]
Pkd1	0.0010	0.0001	2.0E-19	1.0018	163	[28] [G, AS]
Herc2p2	0.0008	0.0001	2.3E-19	1.0012	163	NA
Syne1	0.0018	0.0002	3.0E-18	0.9984	152	[9] [G]
Snrpn	0.0018	0.0002	3.8E-18	1.0020	151	[29, 73][G, AS]
Pde4dip	0.0016	0.0002	1.3E-17	0.9993	146	[30, 74][G, AS]
Golga8c	0.0031	0.0004	4.2E-17	1.0005	141	NA
Sspo	0.0009	0.0001	1.2E-16	1.0003	137	NA
Ankrd36	0.0026	0.0003	1.3E-16	1.0018	137	NA
Tbc1d3	0.0008	0.0001	2.4E-16	1.0026	135	[75][C]
Flj45340	0.0018	0.0002	5.5E-16	1.0007	131	NA
Anapc1	0.0009	0.0001	5.8E-15	0.9990	122	[59, 76][C, AS]
Syne2	0.0012	0.0002	6.2E-15	1.0017	115	[77][C, AS]
Nbpf10	0.0035	0.0005	1.3E-14	0.9992	118	[78] [C, AS]
Muc19	0.0015	0.0002	1.4E-14	1.0001	118	[79][C, AS]
Obscn	0.0006	0.0001	1.5E-14	0.9999	118	[27, 80][G, AS]
Npipl3	0.0019	0.0003	4.1E-14	1.0014	114	NA
Dst	0.0013	0.0002	9.4E-14	0.9997	111	[31, 81][G, AS]

Table 2 (Contd)

Col7a1	0.0011	0.0001	1.4E-13	1.0001	109	[82, 83][C, AS]
Ubr4	0.0011	0.0001	1.4E-13	0.9994	109	[84, 85][C, AS]
Hmcn1	0.0006	0.0001	2.0E-13	0.9975	109	[86][C, AS]
Ryr2	0.0011	0.0001	2.7E-13	0.9974	107	[34, 87][G, AS]
Macf1	0.0011	0.0002	3.1E-13	0.9975	106	[32, 88][G, AS]
Mdn1	0.0006	0.0001	3.5E-13	0.9993	106	NA
Col4a5	0.0008	0.0001	3.5E-13	0.9992	106	[89, 90][C, AS]
Ryr1	0.0007	0.0001	4.2E-13	0.9998	105	[33, 72, 91][G, AS]
Golga6l5	0.0013	0.0002	5.2E-13	1.0021	104	NA
Ryr3	0.0009	0.0001	1.3E-12	0.9962	102	[92][C, AS]
Dnah14	0.0007	0.0001	2.0E-12	0.9990	99	NA
Herc2	0.0006	0.0001	3.1E-12	1.0003	97	[60][AS]
Dnah8	0.0005	0.0001	4.7E-12	0.9997	96	NA
Nomo1	0.0007	0.0001	4.9E-12	0.9996	95	NA
Gpr98	0.0016	0.0002	5.9E-12	0.9948	95	[67][C, AS]
Golga6a	0.0017	0.0002	7.8E-12	1.0009	93	NA

[1]Estimate: exon-survival interaction variance indicator of alternative exon usage;
[2]SE: standard error of the estimate;
[3]P-value AEU: unadjusted P-value of alternative exon usage or exon-dependent association between expression and glioblastoma multiforme survival.
[4]Fold change: fold change in average exon expression per additional survival month;

[5]Exon Count: number of exons in the gene;
[6]Literature: review of studies that reported associations of the gene with cancers:

[G]: reported association of gene with glioblastoma multiforme;

[C]: reported association of gene with cancer other than glioblastoma multiforme.
[AS]: identification of different variants due to alternative splicing (AS) event.

Ttn encodes the protein TTN that is responsible for the passive elasticity of cells. A mutation resulting in an altered TTN was associated with GBM [27]. *Pkd1* was over-expressed during the progression of low-grade to high-grade gliomas [28]. *Syne1* has been associated with increased GBM survival [9]. Under- expression of *Snrnp* was observed in older GBM patients [29]. *Pde4dip* is down-regulated in glioma cell lines treated with dB-cAMP that reduces the invasiveness, proliferation and migratory properties of glioma cells and increases the survival of glioma cells lines [30]. The mutation R4558H in *Obscn* has been associated with GBM [27]. Likewise, a mutation in *Dst* that indirectly regulates the expression of *Otub1* (through regulation of mir-15b has been associated with GBM [31]. Reduced expression of *Macf1* has been observed in glioma cells treated with IL-13 cytotoxin that causes the cells to undergo necrosis. Thus, down-regulation of the expression of *Macf1* is associated with increased GBM survival [32]. *Ryr1* was under-expressed in high-grade gliomas relative to primary (low-grade) gliomas [33]. On the other hand *Ryr2* was over-expressed in invasive GBM cells [34].

Functional and Pathway Analyses of the Multi-Exon Genes Exhibiting Exon-Independent Association with Glioblastoma Multiforme Survival

The list of 2477 genes exhibiting significant evidence of AEU associated with GBM survival was further investigated using functional and pathway analyses. At FDR adjusted P-value < 5.0E-2, 15 KEGG pathways, 87 GO biological processes, and 70 GO molecular functions were enriched. The top 10 pathways, biological processes and molecular functions are summarized in Table 3. Among the 15 KEGG pathways significantly enriched, focal adhesion was the most significant pathway encompassing 86 genes. This result is consistent with many reports of the critical role of focal adhesion and gliomas [35-37]. The extra-cellular matrix- (ECM-) receptor

interaction pathway enrichment detected in this study has been reported in other cancers [38, 39]. The ATP-binding cassette (ABC) transporter pathway has been associated with gliomas [40]. Our finding of small cell lung carcinoma pathways enrichment associated with GBM is consistent with the multiple studies that have identified commonalities among these cancers [41]. The most enriched biological process among the AEU genes associated with GBM survival included regulation of small GTPase mediated signal transduction (RSGST), and neuron development that has been associated with neuroblastoma [42]. The enrichment of biological adhesion confirms our focal adhesion results. Among the top 70 GO molecular functions significantly enriched were: adenyl nucleotide binding, adenyl ribonucleotide binding, ATP binding, nucleotide binding and helicase activity. These related nucleotide binding functions have been associated with GBM [43].

Table 3. Ten most significant KEGG and GO categories enriched among the genes displaying alternative exon usage.

Source	Category	Gene Count[1]	FDR P-value[2]
KEGG Pathway	(hsa04510) focal adhesion	86	3.2E-21
	(hsa04512) ecm-receptor interaction	51	8.5E-20
	(hsa02010) abc transporters	30	2.5E-12
	(hsa04810) regulation of actin cytoskeleton	66	1.7E-07
	(hsa05412) arrhythmogenic right ventricular cardiomyopathy	32	5.9E-06
	(hsa05414) dilated cardiomyopathy	37	1.3E-06
	(hsa04070) phosphatidylinositol signaling system	31	1.2E-05
	(hsa05222) small cell lung cancer	31	3.6E-04
	(hsa05410) hypertrophic cardiomyopathy	32	1.3E-04
	(hsa05200) pathways in cancer	73	3.0E-02
GO Biological Process	(GO:0051056) regulation of small GTPase mediated signal transduction	105	5.0E-25
	(GO:0022610) biological adhesion	197	2.7E-22
	(GO:0007155) cell adhesion	197	2.3E-22
	(GO:0046578) regulation of Ras protein signal transduction	79	5.0E-15

Table 3 (Contd)			
	(GO:0035023) regulation of Rho protein signal transduction	51	1.7E-15
	(GO:0007010) cytoskeleton organization	129	1.3E-15
	(GO:0030029) actin filament-based process	85	2.3E-14
	(GO:0007018) microtubule-based movement	51	2.1E-12
	(GO:0016568) chromatin modification	89	1.9E-12
	(GO:0051276) chromosome organization	132	1.4E-12
GO Molecular Function	(GO:0030554) adenyl nucleotide binding	451	9.9E-59
	(GO:0005524) ATP binding	433	2.2E-59
	(GO:0032559) adenyl ribonucleotide binding	437	2.0E-59
	(GO:0001882) nucleoside binding	456	6.3E-58
	(GO:0001883) purine nucleoside binding	451	1.5E-56
	(GO:0017076) purine nucleotide binding	480	5.2E-44
	(GO:0032555) purine ribonucleotide binding	466	2.9E-44
	(GO:0032553) ribonucleotide binding	466	2.9E-44
	(GO:0000166) nucleotide binding	523	7.4E-39
	(GO:0003774) motor activity	86	1.3E-34

[1]**Gene Count: number of genes that have significant alternative exon usage within category.**

[2]**FDR P-value: False discovery rate adjusted P-value of the hyper-geometric test of category enrichment.**

Multi-Exon Genes Exhibiting Exon-Independent Association with Glioblastoma Multiforme Survival

At unadjusted P-value < 5.0E-4 (equivalent to FDR-adjusted P-value < 5.0E-2), 24 multi-exon genes exhibited exon-independent association with GBM survival (group 2 genes). In other words, there was no evidence of AEU in these genes because the expressions of all the exons were consistently associated with GBM survival and a single general or overall association between the gene and survival can be identified. Table 4 lists the top five multi-exon genes that have the most significant exon-independent association with GBM survival.

Among the 24 multi-exon genes that were associated with GBM survival on a general, exon-independent manner, the five genes that have the lower AEU evidence (AEU unadjusted P-value > 1.0E-3, approximately FDR adjusted P-value > 1.0E-1) are listed in Table 4. The expression of three of these genes increased with increasing survival. Noteworthy was the low number of exons in these genes, relative to the higher number of exons in genes exhibiting evidence of AEU.

Four of five multi-exon genes have been associated to different cancers in studies listed in Table 4 and the remaining gene is uncharacterized (LOC100289627). Sirtuin2 (*Sirt2*) has been associated with GBM while the other three genes golgin subfamily A member 8J (*Golga8j*), semaphorin 3E (*Sema3e*) and SIX homeobox 1 (*SIX1*) were associated with other cancers. Under-expression of *Sirt2* has been reported in glioma cells [44]. This result is also consistent with our findings that higher levels of *Sirt2* were associated with higher GBM hazard. *Golga8j* has been associated with pancreatic cancer and the trend is consistent with our finding of lower GBM survival with higher expression levels of this gene [45].

Table 4. Top 5 multi-exon genes that have significant exon-independent association with glioblastoma multiforme survival.

Gene Symbol	Estimate[1]	SE[2]	Fold Change[3]	P-value[4]	P-val AEU[5]	Exon Count[6]	Literature[7]
Sirt2	0.0337	0.0092	1.0236	3.2E-04	2.5E-03	17	[44][G]
Six1	0.0056	0.0015	1.0039	3.3E-04	2.7E-01	05	[48][C]
Loc 100289627	0.0079	0.0022	1.0055	3.8E-04	4.3E-01	02	NA
Sema3e	-0.0256	0.0066	0.9824	1.3E-04	2.4E-03	18	[46][C]
Golga8j	-0.0536	0.0141	0.9635	1.7E-04	1.1E-03	20	[45][C]

[1]Estimate: change in average exon expression per additional survival month (in log2 units);

[2]SE: standard error of the estimate;

[3]Fold change: fold change in average exon expression per additional survival month;

[4]P-value: unadjusted P-value of the change in average exon expression per additional survival month;

[5]P-value AEU: non-significant (P-value > 1.0E-03) evidence of alternative exon usage;

[6]Exon Count: number of exons in the gene;

[7]Literature: review of studies that reported associations of the gene with cancers;

[G]: reported association of gene with glioblastoma multiforme;
[C]: reported association of gene with cancer other than glioblastoma multiforme.

Sema3e promotes invasiveness and metastatic ability of the cancerous cells [46]. *Sema3e* is associated with many cancers like prostate cancer colon cancer and lung adenocarcinoma [47]. This result is consistent with our findings that higher levels of *Sema3e* were associated with lower GBM survival. The gene *Six1* is associated with lower survival in cancerous cells [48]. This result is inconsistent with our results showing an increase in *Six1* expression associated with an increase in GBM survival.

Single-Exon Genes Associated with Glioblastoma Multiforme Survival

Eight single-exon genes were associated with GBM survival (group 3 genes) at unadjusted P-value < 5.0E-4 (equivalent to FDR-adjusted P-value < 5.0E-2. Table 5 summarizes the results corresponding to these 8 single-exon genes. Among these, three genes had a negative relationship such that lower expression levels were associated with higher survival (Table 5). Four members of the family of small nucleolar RNA CD box (*Snord*) genes were associated with GBM survival and three had a positive association such that higher expression levels were associated with higher survival. These results are consistent with previous work. *Snord* are a type of small nucleolar RNA (SnoRNA) that guide the methylation of rRNAs and snRNAs. These snoRNAs can target other RNAs and are associated with carcinogenesis. Their reduced and dysregulated expression has been associated with progression of many human malignancies [49]. Along with their loss in brain tumorigenesis, snoRNA have also been linked to other cancers such as prostate, breast and lung cancer [49, 50]. In this study, a positive association between the levels of H1 histone family member 0 (*H1f0*) and GBM survival was identified. The expression of *H1f0* was high in breast tumor cells, and decreased when the breast tumor cell lines were reverted back into normal ME cells [51].

Table 5: Results corresponding to 08 single-exon genes associated with glioblastoma multiforme survival (group 3 genes).

Gene symbol	Estimate[1]	SE[2]	Fold Change[3]	P-value[4]	Literature[5]
Hist1h1t	0.0118	0.0024	1.0082	2.5E-06	NA
Snord116-11	0.0101	0.0025	1.0070	9.7E-05	[50][C]
Loc729852	-0.0074	0.0018	0.9949	5.8E-05	NA
Snord123	-0.0087	0.0025	0.9940	4.8E-04	[50][C]
Snord104	0.0067	0.0019	1.0047	4.1E-04	[50][C]
Dkfzp779l1853	-0.0083	0.0023	0.9943	3.9E-04	NA
H1f0	0.0062	0.0017	1.0043	2.3E-04	[51][C]
Snord28	0.0166	0.0044	1.0116	1.8E-04	[50][C]

The table includes group 3 single-exon genes that are associated with glioblastoma multiforme survival.

[1]Estimate: change in gene expression per additional survival month (in log2 units);

[2]SE: standard error of the estimate;

[3]Fold change: fold change in gene expression per additional survival month;

[4]P-value: unadjusted P-value of the change in average exon expression per additional survival month;

[5]Literature: review of studies that reported associations of the gene with cancers;[C]: associated with cancer other than glioblastoma multiforme.

Gene Set Enrichment Analyses of All Genes in Consideration of their Association with Glioblastoma Multiforme Survival

Gene set enrichment analysis considered the level and sign of association between the expression of all the genes studied and GBM survival. At FDR adjusted P-value < 5.0E-2, 94 KEGG pathways, 402 GO biological processes, and 203 GO molecular functions were enriched. Results from the top 10 most significant pathways, biological processes and molecular functions are summarized in Tables 6, 7 and 8. Pathways and GO categories are characterized in GSEA by the number of genes that have a positive or negative association between expression and GBM survival, by the log odds ratio indicating whether the category is more enriched among the genes that have a positive or negative association and the corresponding P-value. Positive (or negative) $\log_e$ odds ratio indicates that the enrichment was higher among the genes with positive (or negative) association with GBM survival. Extreme values indicate higher difference in the enrichment percentages between the positive and negative association groups, meanwhile values close to zero indicate similar enrichment percentages between positive and negative association groups.

Table 6. Ten most significant GO biological processes from the gene set enrichment analysis of the genome.

GO Identifier	GO Biological Process	Over Expressed Gene[1]	Under Expressed Genes[2]	Log Odds Ratio[3]	FDR P-value[4]
GO:0046907	intracellular transport	357	560	-0.7338	3.79E-24
GO:0034613	cellular protein localization	245	433	-0.8490	4.78E-24
GO:0043067	regulation of programmed cell death	351	490	-0.6110	1.68E-15
GO:0016192	vesicle-mediated transport	271	400	-0.6639	1.16E-14
GO:0006629	lipid metabolic process	424	538	-0.5148	1.30E-12
GO:0044265	cellular macromolecule catabolic process	373	485	-0.5379	2.10E-12
GO:0044255	cellular lipid metabolic process	346	457	-0.5528	2.41E-12
GO:0050793	regulation of developmental process	442	549	-0.4932	4.27E-12
GO:0007049	cell cycle	418	522	-0.4978	1.11E-11
GO:0009966	regulation of signal transduction	414	509	-0.4812	9.93E-11

[1]Over Expressed Genes: number of genes that have a positive association between expression and glioblastoma multiforme survival;

[2]Under Expressed Genes: number of genes that have a negative association between expression and glioblastoma multiforme survival;

[3]Log Odds Ratio: indicates whether the category is more enriched among the genes that have a positive association between expression and survival relative to the enrichment among the genes that have a negative association between expression and glioblastoma survival (positive $\log_e$ odds ratio) or vice versa (negative $\log_e$ odds ratio). Extreme values indicate higher difference in the enrichment percentages between the positive and negative association groups meanwhile values close to zero indicate similar enrichment percentages between positive and negative association groups;

[4]FDR P-value: False discovery rate adjusted P-value of the log odds ratio test.

Table 7. Ten most significant GO molecular functions from the gene set enrichment analysis of the genome.

GO Identifier	GO Molecular Function	Over Expressed Gene[1]	Under Expressed Genes[2]	Log Odds Ratio[3]	FDR P-value[4]
GO:0000287	magnesium ion binding	196	300	-0.6962	3.23E-11
GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus containing anhydrides	419	521	-0.4933	5.21E-11
GO:0016462	pyrophosphatase activity	417	520	-0.4962	5.21E-11
GO:0016817	hydrolase activity, acting on acid anhydrides	428	527	-0.4834	9.62E-11
GO:0016773	phosphotransferase activity, alcohol group as acceptor	393	475	-0.4624	5.64E-09
GO:0016301	kinase activity	421	501	-0.4473	5.64E-09
GO:0016788	hydrolase activity, acting on ester bonds	349	429	-0.4781	9.84E-09
GO:0003723	RNA binding	357	437	-0.4741	9.84E-09
GO:0030695	GTPase regulator activity	193	260	-0.5651	4.10E-07
GO:0016874	ligase activity	205	272	-0.5501	4.10E-07

[1]Over Expressed Genes: number of genes that have a positive association between expression and glioblastoma multiforme survival;

[2]Under Expressed Genes: number of genes that have a negative association between expression and glioblastoma multiforme survival;

[3]Log Odds Ratio: indicates whether the category is more enriched among the genes that have a positive association between expression and survival relative to the enrichment among the genes that have a negative association between expression and glioblastoma survival (positive $\log_e$ odds ratio) or vice versa (negative $\log_e$ odds ratio). Extreme values indicate higher difference in the enrichment percentages between the positive and negative association groups meanwhile values close to zero indicate similar enrichment percentages between positive and negative association groups;

[4]FDR P-value: False discovery rate adjusted P-value of the log odds ratio test.

Table 8. Ten most significant KEGG pathways from the gene set enrichment analysis of the genome.

KEGG Identifier	KEGG Pathway	Over Expressed Gene[1]	Under Expressed Genes[2]	Log Odds Ratio[3]	FDR P-value[4]
hsa03010	ribosome	119	16	-2.4779	9.7E-10
hsa00010	glycolysis / gluconeogenesis	57	27	-1.1614	3.6E-04
hsa00190	oxidative phosphorylation	103	39	-0.9392	3.6E-04
hsa05212	pancreatic cancer	54	45	-0.9460	4.7E-04
hsa05130	pathogenic escherichia coli infection	44	41	-1.0575	4.7E-04
hsa00240	pyrimidine metabolism	42	78	-0.8800	5.0E-04
hsa03050	proteasome	33	32	-1.0965	7.2E-04
hsa00280	valine, leucine and isoleucine degradation	20	48	-1.1353	8.5E-04
hsa04662	b cell receptor signaling pathway	34	65	-0.9084	8.5E-04
hsa05223	non-small cell lung cancer	25	52	-0.9922	9.0E-04

[1]Over Expressed Genes: number of genes that have a positive association between expression and glioblastoma multiforme survival;

[2]Under Expressed Genes: number of genes that have a negative association between expression and glioblastoma multiforme survival;

[3]Log Odds Ratio: indicates whether the category is more enriched among the genes that have a positive association between expression and survival relative to the enrichment among the genes that have a negative association between expression and glioblastoma survival (positive $\log_e$ odds ratio) or vice versa (negative $\log_e$ odds ratio). Extreme values indicate higher difference in the enrichment percentages between the positive and negative association groups meanwhile values close to zero indicate similar enrichment percentages between positive and negative association groups;

[4]FDR P-value: False discovery rate adjusted P-value of the log odds ratio test.

Noteworthy was that all top ten results had negative log odds ratio indicating that the categories were more enriched among the genes that have a negative association between expression and survival relative to the enrichment among the genes that have a positive association between expression and GBM survival. Positive log odds ratios were observed for less significant (P-value < 5.0E-2) pathways and categories. The more extreme log odds ratios observed in the GSEA of KEGG pathways indicate higher difference between the enrichment percentages in the positive and negative association groups meanwhile values close to zero in the GSEA of GO categories indicate lower differences in the enrichment percentages between positive and negative association groups.

Among the most differentially enriched pathways (Table 6) were cancer pathways (pancreatic, non-small cell lung). Additional pathways identified in this study that have been associated with gliomas include glycolysis/gluconeogenesis [52] and oxidative phosphorylation [53]. Among the top enriched GO biological processes, lipid metabolism and cell cycle have been associated with glioma [54, 55]. Likewise, several GO molecular functions hydrolase and ligase activities have been have been linked to glioma [56, 57].

Demonstration of Alternative Exon Usage

The identification of patterns of differential exon expression across a gene and comparison against predicted AS models helped to confirm associations between AS and survival. Figures 2.1 to 2.4 depict patterns of exon expression associated with GBM survival and reported AS gene models for three genes among the 36 genes that exhibited the highest significant AEU associated with GBM survival (Table 2) and one gene of biological relevance that have AEU at P-value < 5.0E-4. The four genes depicted in Figures 2.1 to 2.4 are anaphase promoting complex

subunit 1 (*Anapc1*, Figure 2.1), HECT domain and RLD domain containing E3 ubiquitin protein ligase 2 (*Herc2*, Figure 2.2), G-protein coupled receptor 98 (*Gpr98*, Figure 2.3), and epidermal growth factor (*Egf*, Figure 2.4).The parallel alignment of estimated exon expression resulting from our analysis, the moving average trend and the AS prediction from AceView offered *in silico* verification of the identified AEU [3]. The AS models are denoted by lines parallel to the x-axis and identify the corresponding exons. However, no expression values should be assigned to the AS model lines and experimental confirmation of the AEU cases identified in this study is necessary.

Anapc1 is located on human chromosome (HAS) 2 and the function of this gene is associated with transition in the cell cycle from metaphase to anaphase [58]. In agreement with the function, premature truncation of the gene leading to reduced expression of *Anapc1* is associated with cancer development [59]. Six AS models for this gene were found in the alternative transcript variant database ACE View. *Anapc1*exhibited AEU in this study and of the 48 exons analyzed, the expression of 25 exons was associated with GBM survival (Figure 2.1). The AS pattern predicted by our model and highlighted by the moving average trend is supported by AS gene models (Anapc1.d, and Anapc1.e, Figure 2.1). Our model predicted under-expression of the majority of the exons in three gene models (Anapc1.d, Anapc1.e and Anapc1.g). The under expression of exons associated with higher survival predicted by our model and presented in Figure 2.1 are in consistent with previous studies of the relationship between *Anapc1* and cancer [59]. Consistent with the functional analysis, *Anapc1* pertains to enriched GO biological process of cell cycle phase and axonogenesis and the KEGG pathway Ubiquitin mediated proteolysis.

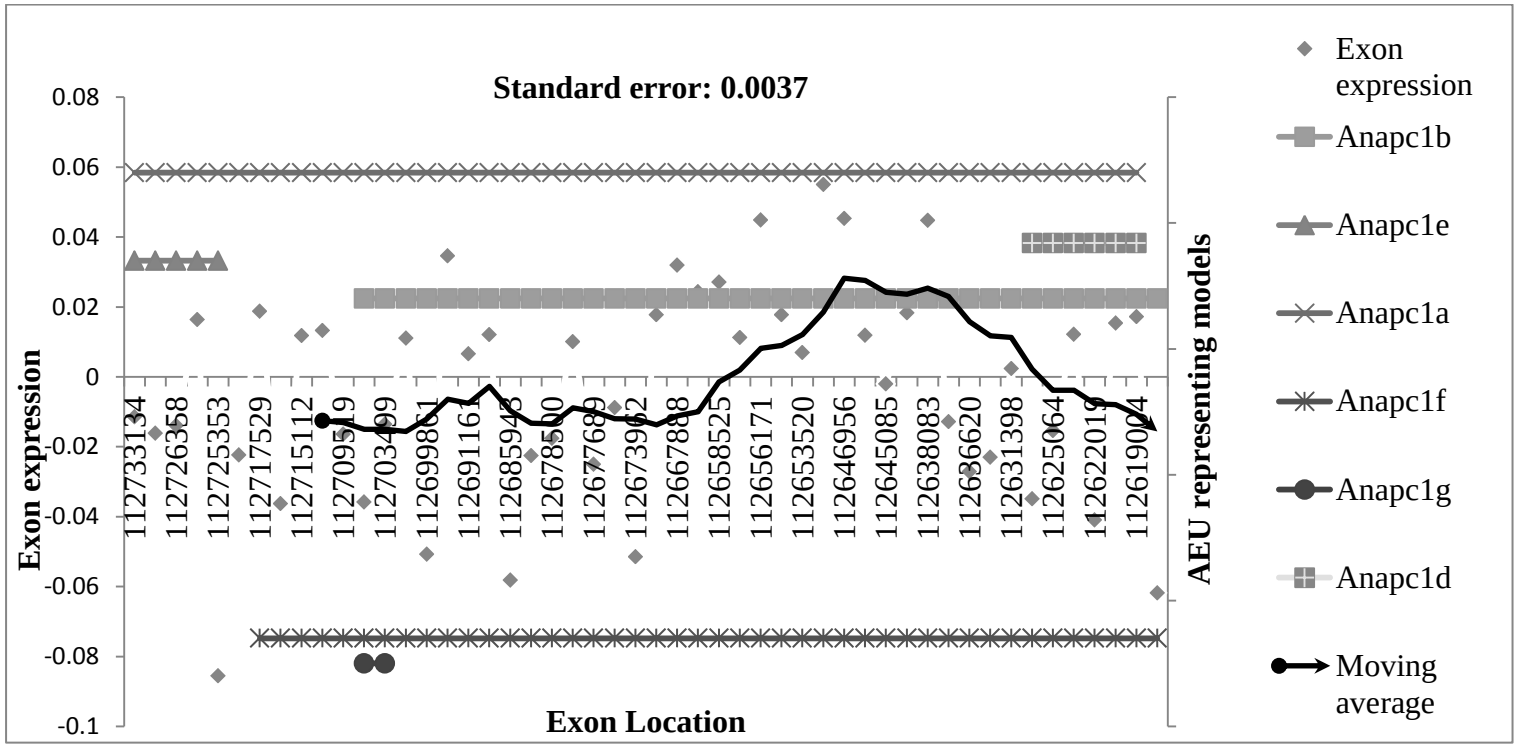
Herc2 is located on HAS 15 and belongs to the ubiquitin ligase family HERC. Various members of this family have high expression in fetal relative to adult brain [60]. *Herc2* in mouse has been associated with neuromuscular disorder, was and has been proposed to be related to neuronal tissues in humans. Also, mutations resulting in under expression of *Herc2* have been related to gastric and colorectal carcinomas [61]. Significant AEU and association between GBM survival and expression were detected on 42 of the 93 exons in *Herc2* (Figure 2.2). Our model predicted exon under-expression that overlap with several AS models (e.g. Herc.q, Herc.g, Herc.j, Herc.t). These trends are consistent with demonstrations that HERC2 depletion restores the breast cancer suppressor BRCA1 [62] and that resulting HERC2 protein formation and cancer [61]. Supporting our GO analyses and enriched categories, *Herc2* belongs to the GO molecular function categories GTPase regulator activity and ion binding, the GO biological process of intracellular transport and protein localization and the KEGG pathway Ubiquitin mediated proteolysis.

Gpr98 is located on HAS 5 and is highly expressed in the central nervous system (CNS) [63]. This gene has been associated with Usher syndrome and Familial Febrile seizures. Usher syndrome is characterized by hearing loss and progressive vision loss, whereas the Febrile Convulsions is a form of seizure effecting children [64, 65]. Studies have r association of *Gpr98* with cancer [66] and also revealed that smaller variants of *Gpr98*, produced due to AS, are associated with increased survival against lymphoblastic leukemia [67]. *Gpr98* exhibited AEU in this study and the expression of approximately 30 exons (out of 90 exons) exhibited significant association with GBM survival (Figure 2.3). Several over-expressed exons detected by our model are consistent with AS gene models including Mass1.b, Mass1.f, Mass1.e, and Mass1.c. Conversely, some under-expressed exons identified in our study are supported by gene models including Mass1d and Mass1g. These results are consistent with previous studies that indicated

association of smaller transcripts of *Gpr98* with cancer survival by inducing apoptosis in cancerous cells [67]. In agreement with our GO analyses, *Gpr98* is affiliated to the enriched GO biological processes of cell adhesion, neuron development and sensory perception of mechanical stimulus. Additionally, *Gpr98* has the GO molecular function of cytoskeletal protein binding and ion binding.

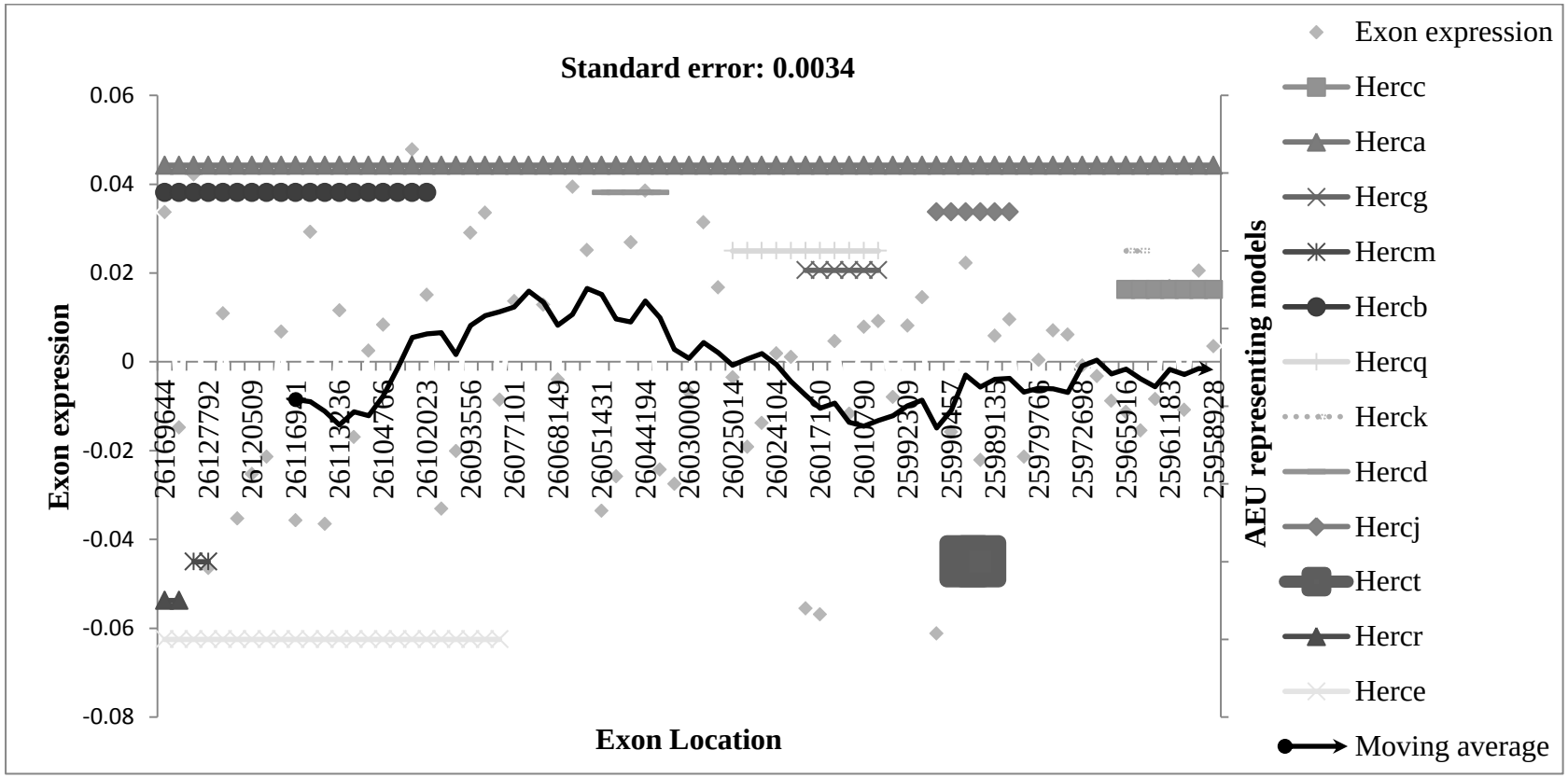
Egf is located on HAS 4 and over-expression of *Egf* has been associated with tumor progression and lower GBM survival [68]. *Egf* exhibited AEU and of the 24 exons analyzed, nine exons had significant associations with GBM survival. Several over-expressed exons detected in our analysis correspond to AS gene models including Egf.j and Egf.h. In accord with the pathway and functional analyses, *Egf* is part of many enriched KEGG pathways including focal adhesion, regulation of actin cytoskeleton, and cancer pathways.

Figure 2.1. *Anapc1* exon expression, moving average, and alternative splicing models.



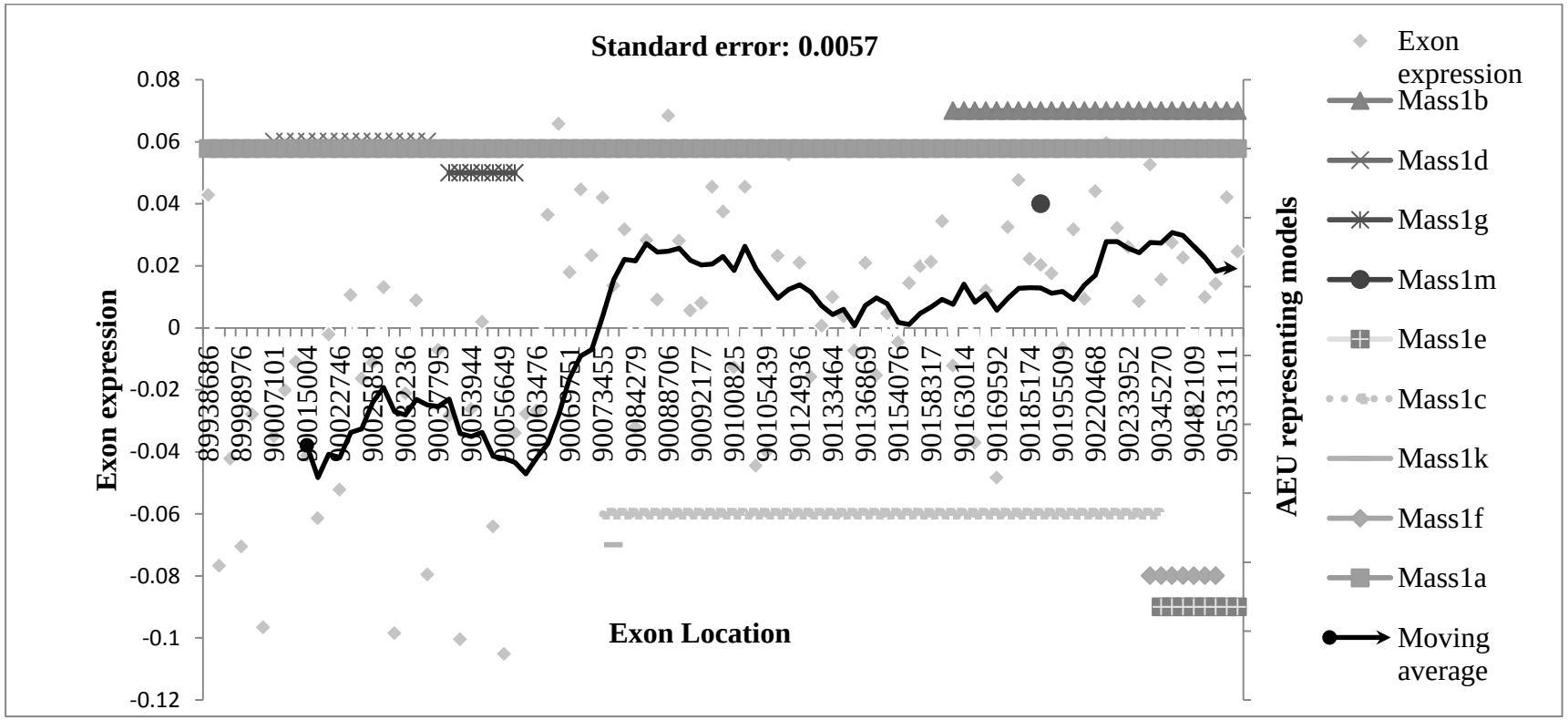
***Anapc1*: anaphase promoting complex subunit 1.**
X-axis: location of exons in the gene (in bp).
Y-axis (left): change in exon expression per additional survival month calculated from the hierarchical model of alternative exon usage.
Full diamond markers: exon expression from the hierarchical model of alternative exon usage is denoted by (Exon expression).
Y-axis (right): indicator of alternative splicing model.
Black and gray lines that have x, triangle, square, circle, +, or no markers: AceView alternative splicing models. Lines denote the location of the exons included in the model. Alternative splicing models do not have inherent expression levels.
Continuous line: moving average pattern of expression based on 10 exons.
Standard Error: standard error of the exon expression estimate.

Figure 2.2. *Herc2* exon expression, moving average, and alternative splicing models.



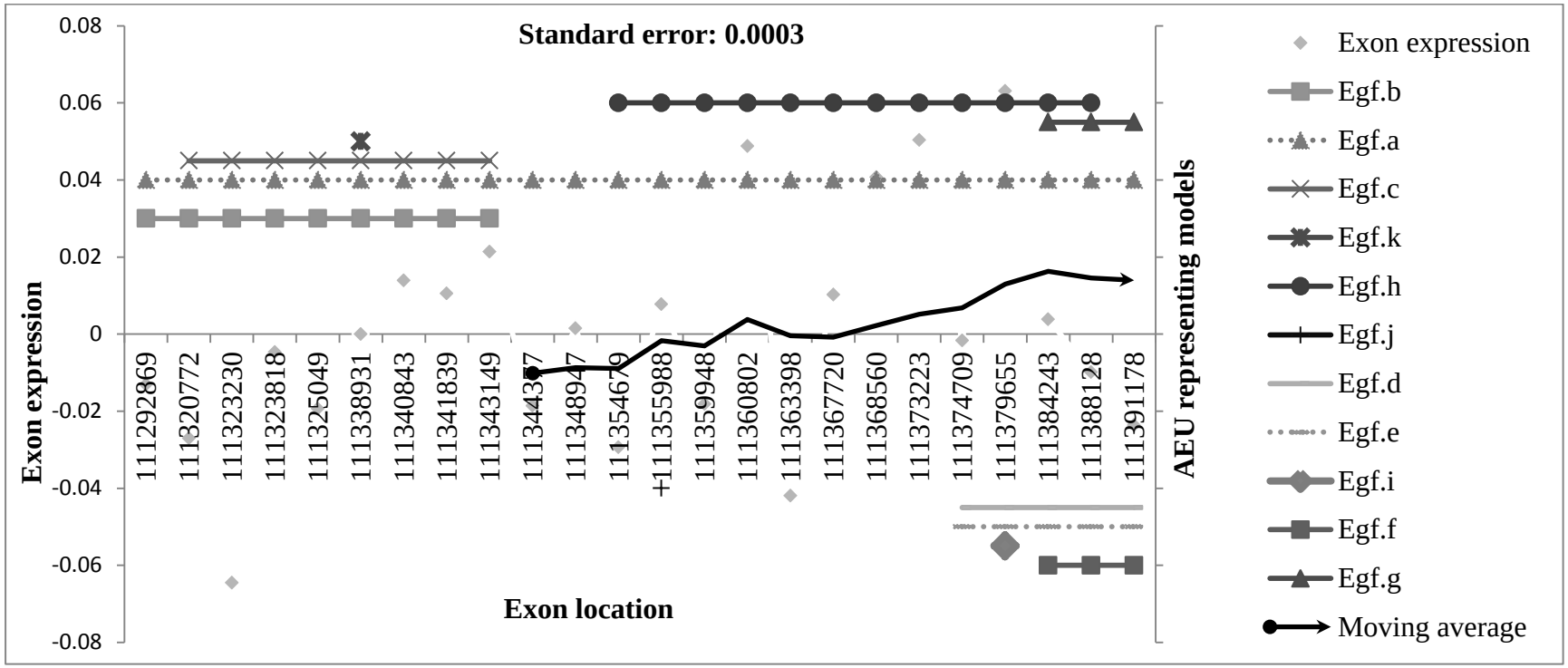
***Herc2*: HECT domain and RLD domain containing E3 ubiquitin protein ligase 2.**
X-axis: location of exons in the gene (in bp).
Y-axis (left): change in exon expression per additional survival month calculated from the hierarchical model of alternative exon usage.
Full diamond markers: exon expression from the hierarchical model of alternative exon usage is denoted by (Exon expression).
Y-axis (right): indicator of alternative splicing model.
Black and gray lines that have x, triangle, square, circle, +, or no markers: AceView alternative splicing models. Lines denote the location of the exons included in the model. Alternative splicing models do not have inherent expression levels.
Continuous line: moving average pattern of expression based on 10 exons.
Standard Error: standard error of the exon expression estimate.

Figure 2.3. *Gpr98* exon expression, moving average, and alternative splicing models.



***Gpr98*: G-protein coupled receptor 98.**
X-axis: location of exons in the gene (in bp).
Y-axis (left): change in exon expression per additional survival month calculated from the hierarchical model of alternative exon usage.
Full diamond markers: exon expression from the hierarchical model of alternative exon usage is denoted by (Exon expression).
Y-axis (right): indicator of alternative splicing model.
Black and gray lines that have x, triangle, square, circle, +, or no markers: AceView alternative splicing models. Lines denote the location of the exons included in the model. Alternative splicing models do not have inherent expression levels.
Continuous line: moving average pattern of expression based on 10 exons.
Standard Error: standard error of the exon expression estimate.

Figure 2.4. *Egf* exon expression, moving average, and alternative splicing models.



***Egf*: epidermal growth factor.**
X-axis: location of exons in the gene (in bp).
Y-axis (left): change in exon expression per additional survival month calculated from the hierarchical model of alternative exon usage.
Full diamond markers: exon expression from the hierarchical model of alternative exon usage is denoted by (Exon expression).
Y-axis (right): indicator of alternative splicing model.
Black and gray lines that have x, triangle, square, circle, +, or no markers: AceView alternative splicing models. Lines denote the location of the exons included in the model. Alternative splicing models do not have inherent expression levels.
Continuous line: moving average pattern of expression based on 10 exons.
Standard Error: standard error of the exon expression estimate.

Validation

The average and range of R^2 values from the application of training data set estimates on the training and validation data set were 0.8 and 0.7, respectively. These results indicate that the AEU events and genes associated with GBM survival detected and characterized in the training data set were confirmed in the independent validation data set.

Further Studies

Extensions to the hierarchical model proposed in this study to identify AEU can be considered. First, the model can incorporate information of the mapping of the exons to the gene. In addition, the distance between the exons can be accommodated on the variance-covariance matrix. This would allow modeling of potentially higher dependencies between proximal exons relative to distant exons. Second, the model can incorporate information on different splicing scenarios [3].

In this study, the vast majority of the exons within a gene mapped to one strand and few exons mapped to the other strand. Thus, AEU was studied among the exons that mapped to the most frequent strand. When sufficient information on both strands within a gene is available, our model allows the consideration of information across strands. This model would allow the study of sense-antisense gene overlap and its impact on AS and regulation of gene expression following the work of Sorana Morrissy et al. Their work suggested an antisense transcription-mediated mechanism of splicing regulation in human cells [69].

Conclusions

In conclusion, AEU is a complex process and thus the detection and characterization of AEU associated with survival is challenging. The hierarchical model developed in this study allowed simultaneously, the detection of differential expression of exons within a gene and differentially expressed genes associated with survival. From a total of 25,403 genes investigated, 2477 multi-exon and 13 single exon genes were associated with GBM. Most of the significant genes detected by the model have been previously associated to GBM (27.78%) or other type of cancer (36.11%). The AEU events detected for several genes (*Egf*, *Herc2*, *Gpr98*, *Anapc1*) were consistent with AS models in AceView. The hierarchical model can be applied to other cancer types and to indicators other than survival.

References

1. Sakabe NJ, Vibranovski MD, de Souza SJ: **A bioinformatics analysis of alternative exon usage in human genes coding for extracellular matrix proteins.** Genet Mol Res 2004, **3**(4):532-544.

2. Barash Y, Blencowe BJ, Frey BJ: **Model-based detection of alternative splicing signals.** Bioinformatics 2010, **26**(12):i325-33.

3. Laderas TG, Walter NA, Mooney M, Vartanian K, Darakjian P, Buck K, Harrington CA, Belknap J, Hitzemann R, McWeeney SK: **Computational detection of alternative exon usage.** Front Neurosci 2011, **5**:69.

4. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** Science 2003, **302**(5653):2141-2144.

5. Kim H, Klein R, Majewski J, Ott J: **Estimating rates of alternative splicing in mammals and invertebrates.** Nat Genet 2004, **36**(9):915-6; author reply 916-7.

6. Ramskold D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.** PLoS Comput Biol 2009, **5**(12):e1000598.

7. Yeo G, Holste D, Kreiman G, Burge CB: **Variation in alternative splicing across human tissues.** Genome Biol 2004, **5**(10):R74.

8. 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.

9. Serao NV, Delfino KR, Southey BR, Beever JE, Rodriguez-Zas SL: **Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival.** BMC Med Genomics 2011, **4**:49.

10. Delfino KR, Serao NV, Southey BR, Rodriguez-Zas SL: **Therapy-, gender- and race-specific microRNA markers, target genes and networks related to glioblastoma recurrence and survival.** Cancer Genomics Proteomics 2011, **8**(4):173-183.

11. Lo HW, Zhu H, Cao X, Aldrich A, Ali-Osman F: **A novel splice variant of GLI1 that promotes glioblastoma cell migration and invasion.** Cancer Res 2009, **69**(17):6790-6798.

12. Johnson DR, O'Neill BP: **Glioblastoma survival in the United States before and during the temozolomide era.** J Neurooncol 2012, **107**(2):359-364.

13. Krex D, Klink B, Hartmann C, von Deimling A, Pietsch T, Simon M, Sabel M, Steinbach JP, Heese O, Reifenberger G, Weller M, Schackert G, German Glioma Network: **Long-term survival with glioblastoma multiforme.** Brain 2007, **130**(Pt 10):2596-2606.

14. Lamborn KR, Chang SM, Prados MD: **Prognostic factors for survival of patients with glioblastoma: recursive partitioning analysis.** Neuro Oncol 2004, **6**(3):227-235.

15. Su WL, Modrek B, GuhaThakurta D, Edwards S, Shah JK, Kulkarni AV, Russell A, Schadt EE, Johnson JM, Castle JC: **Exon and junction microarrays detect widespread mouse strain- and sex-bias expression differences.** BMC Genomics 2008, **9**:273.

16. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP: **FIRMA: a method for detection of alternative splicing from exon array data.** Bioinformatics 2008, **24**(15):1707-1714.

17. Zheng H, Hang X, Zhu J, Qian M, Qu W, Zhang C, Deng M: **REMAS: a new regression model to identify alternative splicing events from exon array data.** BMC Bioinformatics 2009, **10 Suppl 1**:S18.

18. Cline MS, Blume J, Cawley S, Clark TA, Hu JS, Lu G, Salomonis N, Wang H, Williams A: **ANOSVA: a statistical method for detecting splice variation from expression data.** Bioinformatics 2005, **21 Suppl 1**:i107-15.

19. Maronna RA, Martin RD, Yohai VJ: *Robust Statistics: Theory And Methods:* illustrated ed. J. Wiley; 2006.

20. Klipper-Aurbach Y, Wasserman M, Braunspiegel-Weintrob N, Borstein D, Peleg S, Assa S, Karp M, Benjamini Y, Hochberg Y, Laron Z: **Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus.** Med Hypotheses 1995, **45**(5):486-490.

21. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** Nucleic Acids Res 2009, **37**(1):1-13.

22. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** Nat Protoc 2009, **4**(1):44-57.

23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** Proc Natl Acad Sci U S A 2005, **102**(43):15545-15550.

24. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tarraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, Garcia F, Marba M, Montaner D, Dopazo J: **Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling.** Nucleic Acids Res 2010, **38**(Web Server issue):W210-3.

25. Ji H, Wong WH: **TileMap: create chromosomal map of tiling array hybridizations.** Bioinformatics 2005, **21**(18):3629-3636.

26. Olson KM, Vanraden PM, Tooker ME, Cooper TA: **Differences among methods to validate genomic evaluations for dairy cattle.** J Dairy Sci 2011, **94**(5):2613-2620.

27. Balakrishnan A, Bleeker FE, Lamba S, Rodolfo M, Daniotti M, Scarpa A, van Tilborg AA, Leenstra S, Zanon C, Bardelli A: **Novel somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma.** Cancer Res 2007, **67**(8):3545-3550.

28. Ma Y, Yuan RQ, Fan S, Hu C, Goldberg ID, Laterra JJ, Rosen EM: **Identification of genes that modulate sensitivity of U373MG glioblastoma cells to cis-platinum.** Anticancer Drugs 2006, **17**(7):733-751.

29. Korshunov A, Sycheva R, Golanov A: **Genetically distinct and clinically relevant subtypes of glioblastoma defined by array-based comparative genomic hybridization (array-CGH).** Acta Neuropathol 2006, **111**(5):465-474.

30. Moreno MJ, Ball M, Andrade MF, McDermid A, Stanimirovic DB: **Insulin-like growth factor binding protein-4 (IGFBP-4) is a novel anti-angiogenic and anti-tumorigenic mediator secreted by dibutyryl cyclic AMP (dB-cAMP)-differentiated glioblastoma cells.** Glia 2006, **53**(8):845-857.

31. Dong H, Luo L, Hong S, Siu H, Xiao Y, Jin L, Chen R, Xiong M: **Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma.** BMC Syst Biol 2010, **4**:163.

32. Han J, Yang L, Puri RK: **Analysis of target genes induced by IL-13 cytotoxin in human glioblastoma cells.** J Neurooncol 2005, **72**(1):35-46.

33. van den Boom J, Wolter M, Kuick R, Misek DE, Youkilis AS, Wechsler DS, Sommer C, Reifenberger G, Hanash SM: **Characterization of gene expression profiles associated with glioma progression using oligonucleotide-based microarray analysis and real-time reverse transcription-polymerase chain reaction.** Am J Pathol 2003, **163**(3):1033-1043.

34. Hoelzinger DB, Mariani L, Weis J, Woyke T, Berens TJ, McDonough WS, Sloan A, Coons SW, Berens ME: **Gene expression profile of glioblastoma multiforme invasive phenotype points to new therapeutic targets.** Neoplasia 2005, **7**(1):7-16.

35. Francescone R, Scully S, Bentley B, Yan W, Taylor SL, Oh D, Moral L, Shao R: **Glioblastoma-Derived Tumor Cells Induce Vasculogenic Mimicry through Flk-1 Activation.** J Biol Chem 2012, .

36. Garzon-Muvdi T, Schiapparelli P, ap Rhys C, Guerrero-Cazares H, Smith C, Kim DH, Kone L, Farber H, Lee DY, An SS, Levchenko A, Quinones-Hinojosa A: **Regulation of brain tumor dispersal by NKCC1 through a novel role in focal adhesion regulation.** PLoS Biol 2012, **10**(5):e1001320.

37. Swiatek-Machado K, Mieczkowski J, Ellert-Miklaszewska A, Swierk P, Fokt I, Szymanski S, Skora S, Szeja W, Grynkiewicz G, Lesyng B, Priebe W, Kaminska B: **Novel small molecular inhibitors disrupt the JAK/STAT3 and FAK signaling pathways and exhibit a potent antitumor activity in glioma cells.** Cancer Biol Ther 2012, **13**(8):657-670.

38. Jiang ZQ, Gui SB, Zhang YZ: **Differential gene expression by fiber-optic beadarray and pathway in adrenocorticotrophin-secreting pituitary adenomas.** Chin Med J (Engl) 2010, **123**(23):3455-3461.

39. Qiu X, Guo S, Wu H, Chen J, Zhou Q: **Identification of Wnt pathway, uPA, PAI-1, MT1-MMP, S100A4 and CXCR4 associated with enhanced metastasis of human large cell lung cancer by DNA microarray.** Minerva Med 2012, **103**(3):151-164.

40. Kievit FM, Wang FY, Fang C, Mok H, Wang K, Silber JR, Ellenbogen RG, Zhang M: **Doxorubicin loaded iron oxide nanoparticles overcome multidrug resistance in cancer in vitro.** J Control Release 2011, **152**(1):76-83.

41. Zheng M, Morgan-Lappe SE, Yang J, Bockbrader KM, Pamarthy D, Thomas D, Fesik SW, Sun Y: **Growth inhibition and radiosensitization of glioblastoma and lung cancer cells by small interfering RNA silencing of tumor necrosis factor receptor-associated factor 2.** Cancer Res 2008, **68**(18):7570-7578.

42. Serra A, Haberle B, Konig IR, Kappler R, Suttorp M, Schackert HK, Roesner D, Fitze G: **Rare occurrence of PHOX2b mutations in sporadic neuroblastomas.** J Pediatr Hematol Oncol 2008, **30**(10):728-732.

43. Lymbouridou R, Soufla G, Chatzinikola AM, Vakis A, Spandidos DA: **Down-regulation of K-ras and H-ras in human brain gliomas.** Eur J Cancer 2009, **45**(7):1294-1303.

44. Haigis MC, Sinclair DA: **Mammalian sirtuins: biological insights and disease relevance.** Annu Rev Pathol 2010, **5**:253-295.

45. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ et al: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** Science 2008, **321**(5897):1801-1806.

46. Casazza A, Kigel B, Maione F, Capparuccia L, Kessler O, Giraudo E, Mazzone M, Neufeld G, Tamagnone L: **Tumour growth inhibition and anti-metastatic activity of a mutated furin-resistant Semaphorin 3E isoform.** EMBO Mol Med 2012, **4**(3):234-250.

47. Blanc V, Nariculam J, Munson P, Freeman A, Klocker H, Masters J, Williamson M: **A role for class 3 semaphorins in prostate cancer.** Prostate 2011, **71**(6):649-658.

48. Qamar L, Deitsch E, Patrick AN, Post MD, Spillman MA, Iwanaga R, Thorburn A, Ford HL, Behbakht K: **Specificity and prognostic validation of a polyclonal antibody to detect Six1 homeoprotein in ovarian cancer.** Gynecol Oncol 2012, .

49. Askarian-Amiri ME, Crawford J, French JD, Smart CE, Smith MA, Clark MB, Ru K, Mercer TR, Thompson ER, Lakhani SR, Vargas AC, Campbell IG, Brown MA, Dinger ME, Mattick JS: **SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer.** RNA 2011, **17**(5):878-891.

50. Liao J, Yu L, Mei Y, Guarnera M, Shen J, Li R, Liu Z, Jiang F: **Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer.** 2010, **9**.

51. Klein A, Guhl E, Zollinger R, Tzeng YJ, Wessel R, Hummel M, Graessmann M, Graessmann A: **Gene expression profiling: cell cycle deregulation and aneuploidy do not cause breast cancer formation in WAP-SVT/t transgenic animals.** J Mol Med (Berl) 2005, **83**(5):362-376.

52. Oudard S, Boitier E, Miccoli L, Rousset S, Dutrillaux B, Poupon MF: **Gliomas are driven by glycolysis: putative roles of hexokinase, oxidative phosphorylation and mitochondrial ultrastructure.** Anticancer Res 1997, **17**(3C):1903-1911.

53. Clarion L, Schindler M, de Weille J, Lolmede K, Laroche-Clary A, Uro-Coste E, Robert J, Mersel M, Bakalara N: **7beta-Hydroxycholesterol-induced energy stress leads to sequential opposing signaling responses and to death of C6 glioblastoma cells.** Biochem Pharmacol 2012, **83**(1):37-46.

54. Laurenti G, Benedetti E, D'Angelo B, Cristiano L, Cinque B, Raysi S, Alecci M, Ceru MP, Cifone MG, Galzio R, Giordano A, Cimini A: **Hypoxia induces peroxisome proliferator-activated receptor alpha (PPARalpha) and lipid metabolism peroxisomal enzymes in human glioblastoma cells.** J Cell Biochem 2011, **112**(12):3891-3901.

55. Liu S, Yin F, Fan W, Wang S, Guo XR, Zhang JN, Tian Z, Fan M: **Over-expression of BMPR-IB reduces the malignancy of glioblastoma cells by upregulation of p21 and p27Kip1.** J Exp Clin Cancer Res 2012, **31**(1):52.

56. Lenman A, Fowler CJ: **Interaction of ligands for the peroxisome proliferator-activated receptor gamma with the endocannabinoid system.** Br J Pharmacol 2007, **151**(8):1343-1351.

57. Jiang X, Xing H, Kim TM, Jung Y, Huang W, Yang HW, Song S, Park PJ, Carroll RS, Johnson MD: **Numb regulates glioma stem cell fate and growth by altering epidermal growth factor receptor and skp1-cullin-f-box ubiquitin ligase activity.** Stem Cells 2012, **30**(7):1313-1326.

58. Jorgensen PM, Graslund S, Betz R, Stahl S, Larsson C, Hoog C: **Characterisation of the human APC1, the largest subunit of the anaphase-promoting complex.** Gene 2001, **262**(1-2):51-59.

59. He ML, Chen Y, Chen Q, He Y, Zhao J, Wang J, Yang H, Kung HF: **Multiple gene dysfunctions lead to high cancer-susceptibility: evidences from a whole-exome sequencing study.** Am J Cancer Res 2011, **1**(4):562-573.

60. Hochrainer K, Mayer H, Baranyi U, Binder B, Lipp J, Kroismayr R: **The human HERC family of ubiquitin ligases: novel members, genomic organization, expression profiling, and evolutionary aspects.** Genomics 2005, **85**(2):153-164.

61. Xu L, Drachenberg C, Burke A: **Intimal IgM lambda paraprotein deposition in myocardial arteries resulting in acute myocardial infarction and sudden death.** Pathology 2011, **43**(7):732-734.

62. Wu W, Sato K, Koike A, Nishikawa H, Koizumi H, Venkitaraman AR, Ohta T: **HERC2 is an E3 ligase that targets BRCA1 for degradation.** Cancer Res 2010, **70**(15):6384-6392.

63. Piro RM, Molineris I, Ala U, Di Cunto F: **Evaluation of candidate genes from orphan FEB and GEFS+ loci by analysis of human brain gene expression atlases.** PLoS One 2011, **6**(8):e23149.

64. Millan JM, Aller E, Jaijo T, Blanco-Kelly F, Gimenez-Pardo A, Ayuso C: **An update on the genetics of usher syndrome.** J Ophthalmol 2011, **2011**:417217.

65. Nakayama J, Fu YH, Clark AM, Nakahara S, Hamano K, Iwasaki N, Matsui A, Arinami T, Ptacek LJ: **A nonsense mutation of the MASS1 gene in a family with febrile and afebrile seizures.** Ann Neurol 2002, **52**(5):654-657.

66. Nagayama K, Kohno T, Sato M, Arai Y, Minna JD, Yokota J: **Homozygous deletion scanning of the lung cancer genome at a 100-kb resolution.** Genes Chromosomes Cancer 2007, **46**(11):1000-1010.

67. Rainer J, Lelong J, Bindreither D, Mantinger C, Ploner C, Geley S, Kofler R: **Research resource: transcriptional response to glucocorticoids in childhood acute lymphoblastic leukemia.** Mol Endocrinol 2012, **26**(1):178-193.

68. Sjostrom S, Andersson U, Liu Y, Brannstrom T, Broholm H, Johansen C, Collatz-Laier H, Henriksson R, Bondy M, Melin B: **Genetic variations in EGF and EGFR and glioblastoma outcome.** Neuro Oncol 2010, **12**(8):815-821.

69. Morrissy AS, Griffith M, Marra MA: **Extensive relationship between antisense transcription and alternative splicing in the human genome.** Genome Res 2011, **21**(8):1203-1212.

70. Gewandter JS, Bambara RA, O'Reilly MA: **The RNA surveillance protein SMG1 activates p53 in response to DNA double-strand breaks but not exogenously oxidized mRNA.** Cell Cycle 2011, **10**(15):2561-2567.

71. Donner K, Sandbacka M, Lehtokari VL, Wallgren-Pettersson C, Pelin K: **Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts.** Eur J Hum Genet 2004, **12**(9):744-751.

72. Difilippantonio S, Chen Y, Pietas A, Schluns K, Pacyna-Gengelbach M, Deutschmann N, Padilla-Nash HM, Ried T, Petersen I: **Gene expression profiles in human non-small and small-cell lung cancers.** Eur J Cancer 2003, **39**(13):1936-1947.

73. Dittrich B, Buiting K, Korn B, Rickard S, Buxton J, Saitoh S, Nicholls RD, Poustka A, Winterpacht A, Zabel B, Horsthemke B: **Imprint switching on human chromosome 15 may involve alternative transcripts of the SNRPN gene.** Nat Genet 1996, **14**(2):163-170.

74. 't Hoen PA, Hirsch M, de Meijer EJ, de Menezes RX, van Ommen GJ, den Dunnen JT: **mRNA degradation controls differentiation state-dependent differences in transcript and splice variant abundance.** Nucleic Acids Res 2011, **39**(2):556-566.

75. Hodzic D, Kong C, Wainszelbaum MJ, Charron AJ, Su X, Stahl PD: **TBC1D3, a hominoid oncoprotein, is encoded by a cluster of paralogues located on chromosome 17q12.** Genomics 2006, **88**(6):731-736.

76. Gamsiz ED, Ouyang Q, Schmidt M, Nagpal S, Morrow EM: **Genome-wide transcriptome analysis in murine neural retina using high-throughput RNA sequencing.** Genomics 2012, **99**(1):44-51.

77. Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh C, Gervais-Bird J, Lapointe E, Froehlich U, Durand M, Gendron D, Brosseau JP, Thibault P, Lucier JF, Tremblay K, Prinos P, Wellinger RJ, Chabot B, Rancourt C, Elela SA: **Identification of alternative splicing markers for breast cancer.** Cancer Res 2008, **68**(22):9525-9531.

78. Vandepoele K, Andries V, Van Roy N, Staes K, Vandesompele J, Laureys G, De Smet E, Berx G, Speleman F, van Roy F: **A constitutional translocation t(1;17)(p36.2;q11.2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes.** PLoS One 2008, **3**(5):e2207.

79. Das B, Cash MN, Hand AR, Shivazad A, Culp DJ: **Expression of Muc19/Smgc gene products during murine sublingual gland development: cytodifferentiation and maturation of salivary mucous cells.** J Histochem Cytochem 2009, **57**(4):383-396.

80. Kim JH, Park BL, Pasaje CF, Kim Y, Bae JS, Park JS, Uh ST, Kim YH, Kim MK, Choi IS, Cho SH, Choi BW, Koh I, Park CS, Shin HD: **Contribution of the OBSCN Nonsynonymous Variants to Aspirin Exacerbated Respiratory Disease Susceptibility in Korean Population.** DNA Cell Biol 2012, .

81. Boutz PL, Stoilov P, Li Q, Lin CH, Chawla G, Ostrow K, Shiue L, Ares M,Jr, Black DL: **A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons.** Genes Dev 2007, **21**(13):1636-1652.

82. Kita Y, Mimori K, Tanaka F, Matsumoto T, Haraguchi N, Ishikawa K, Matsuzaki S, Fukuyoshi Y, Inoue H, Natsugoe S, Aikou T, Mori M: **Clinical significance of LAMB3 and COL7A1 mRNA in esophageal squamous cell carcinoma.** Eur J Surg Oncol 2009, **35**(1):52-58.

83. Wessagowit V, Nalla VK, Rogan PK, McGrath JA: **Normal and abnormal mechanisms of gene splicing and relevance to inherited skin diseases.** J Dermatol Sci 2005, **40**(2):73-84.

84. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, Davies H, Edkins S, Hardy C, Latimer C, Teague J, Andrews J, Barthorpe S, Beare D, Buck G, Campbell PJ, Forbes S, Jia M, Jones D, Knott H, Kok CY, Lau KW, Leroy C, Lin ML, McBride DJ, Maddison M, Maguire S, McLay K, Menzies A, Mironenko T, Mulderrig L, Mudie L, O'Meara S, Pleasance E, Rajasingham A, Shepherd R, Smith R, Stebbings L, Stephens P, Tang G, Tarpey PS, Turrell K, Dykema KJ, Khoo SK, Petillo D, Wondergem B, Anema J, Kahnoski RJ, Teh BT, Stratton MR, Futreal PA: **Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes.** Nature 2010, **463**(7279):360-363.

85. Fehlbaum-Beurdeley P, Jarrige-Le Prado AC, Pallares D, Carriere J, Guihal C, Soucaille C, Rouet F, Drouin D, Sol O, Jordan H, Wu D, Lei L, Einstein R, Schweighoffer F, Bracco L: **Toward an Alzheimer's disease diagnosis via high-resolution blood gene expression.** Alzheimers Dement 2010, **6**(1):25-38.

86. Arne G, Kristiansson E, Nerman O, Kindblom LG, Ahlman H, Nilsson B, Nilsson O: **Expression profiling of GIST: CD133 is associated with KIT exon 11 mutations, gastric location and poor prognosis.** Int J Cancer 2011, **129**(5):1149-1161.

87. George CH, Rogers SA, Bertrand BM, Tunwell RE, Thomas NL, Steele DS, Cox EV, Pepper C, Hazeel CJ, Claycomb WC, Lai FA: **Alternative splicing of ryanodine receptors modulates cardiomyocyte Ca2+ signaling and susceptibility to apoptosis.** Circ Res 2007, **100**(6):874-883.

88. Misquitta-Ali CM, Cheng E, O'Hanlon D, Liu N, McGlade CJ, Tsao MS, Blencowe BJ: **Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer.** Mol Cell Biol 2011, **31**(1):138-150.

89. Gorlov IP, Byun J, Gorlova OY, Aparicio AM, Efstathiou E, Logothetis CJ: **Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data.** BMC Med Genomics 2009, **2**:48.

90. Lemmink HH, Kluijtmans LA, Brunner HG, Schroder CH, Knebelmann B, Jelinkova E, van Oost BA, Monnens LA, Smeets HJ: **Aberrant splicing of the COL4A5 gene in patients with Alport syndrome.** Hum Mol Genet 1994, **3**(2):317-322.

91. Kimura T, Lueck JD, Harvey PJ, Pace SM, Ikemoto N, Casarotto MG, Dirksen RT, Dulhunty AF: **Alternative splicing of RyR1 alters the efficacy of skeletal EC coupling.** Cell Calcium 2009, **45**(3):264-274.

92. Zhang L, Liu Y, Song F, Zheng H, Hu L, Lu H, Liu P, Hao X, Zhang W, Chen K: **Functional SNP in the microRNA-367 binding site in the 3'UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification.** Proc Natl Acad Sci U S A 2011, **108**(33):13653-13658.