BIOINFORMATIC ANALYSIS OF CIS-ENCODED ANTISENSE TRANSCRIPTION

by

Anca Sorana Morrissy

B.Sc., Simon Fraser University, 2002

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2010

Abstract

A key first step in understanding cellular processes is a quantitative and comprehensive measurement of gene expression profiles. The scale and complexity of the mammalian transcriptome is a significant challenge to efforts aiming to identify the complete set of expressed transcripts. Specifically, detection of low-abundance sequences, such as antisense transcripts, has historically been difficult to achieve using EST libraries, microarrays, or tag sequencing methods. Antisense transcripts are expressed from the opposite strand of a partner gene, and in some cases can regulate the processing of the sense transcript, highlighting their biological relevance. Recently, efficient profiling of low-frequency transcripts was made possible with the advent of next generation sequencing platforms. Thus, a major goal of my thesis was to assess the prevalence of antisense transcripts using Tag-seq, a tag sequencing method modified to take advantage of the Illumina sequencing platform. The increase in sampling depth provided by Tag-seq resulted in significantly improved detection of low abundance antisense transcripts, and allowed accurate measurements of their differential expression across normal and cancerous states.

While antisense transcription is known to regulate sense transcript processing at a small number of loci, no genome wide assessments of this regulatory interaction exist. I addressed this knowledge gap using Affymetrix exon arrays, and found a significant correlation between antisense transcription and alternative splicing in normal human cells. Further exploring the biological relevance of antisense-correlated splicing events in human disease, I found that these events could be used to identify clinically distinct subtypes of cancer. Together, the findings in this thesis provide a new foundation for the investigation of antisense transcripts in the regulation of alternative transcript processing, and open new avenues of research into understanding the molecular heterogeneity of human cancers.

Preface

In conjunction with my supervisor, Dr. Marco Marra, I was involved in the conceptualization and design of the research activities described in this thesis. In particular, I was responsible for the design and implementation of the computational experiments, data analysis, and generation of the tables, figures, and text in this thesis. **Chapter 2** represents a collaborative research effort that involved the input of additional authors, as described here and at the beginning of that chapter. Thomas Zhang, Helen McDonald, Yonjun Zhao., and Martin Hirst were involved in the design of the Tag-seq method, and in the creation of Tag-seq and LongSAGE libraries analyzed (sections 2.2.1, 2.4.1, 2.4.2). Steven Jones and Marco Marra supervised the project and contributed design concepts and comments throughout. Allen Delaney designed the data filtering algorithm, and conducted inter-plaform correlation measures (sections 2.2.2, 2.4.2, 2.4.3, 2.2.4.1, 2.2.4.2, 2.4.6). Scott Diguistini processed RNA-seq data. Ryan Morin wrote scripts used in section 2.4.7, was involved in original manuscript planning, and contributed text to the manuscript. I performed the remaining computational experiments in the text, designed and performed the cancer-related analyses (sections 2.2.2-2.2.8 and 2.4.4, 2.4.5, 2.4.7-2.4.9), created figures and tables describing the results, and wrote the majority of the manuscript. The work described in **Chapter 3** was entirely performed and written by me, with the exception of microarray data pre-processing, which was done by Malachi Griffith (section 3.4.3.1), who also contributed microarray analysis concepts. The work described in **Chapter 4** was entirely conducted by me..

Table of Contents

.

List of Tables

List of Figures

Acknowledgements

My most sincere gratitude goes to my supervisor, Dr. Marco Marra, a steadfast supporter and guide throughout the years of my PhD. His exceptional work ethic, leadership skills, personal integrity, and scientific prowess form a personal and scientific standard to which I will always aspire. To the members of my thesis advisory committee, Drs. Pamela Hoodless, Dixie Mager, and Steven Jones, I am indebted for timely and practical advice, and for their astute scientific perspectives. Their investment of time and expertise was invaluable to the success of my work.

I am grateful to have been part of the Marra lab, and would like to thank Malachi Griffith and Ryan Morin in particular for their collaborative support that contributed to this thesis, and Olena Morozova and Rodrigo Goya for insightful discussions. Besides intellectual and technical advice, my fellow labmates have been a great source of positive attitude and fun. I extend my thanks to the staff and scientists at the BC Cancer Agency's Genome Sciences Center (GSC). The success of my work was in large part due to the exceptionally creative, energetic, knowledgeable and focused co-workers I have had the privilege to interact with, particularly: Obi Griffith, Monica Sleumer, Claire Hou, Erin Pleasance, Mikhail (Misha) Bilenky, Yaron Butterfield, Hye Jung (Elizabeth) Chun, Allen Delaney, Noushin Farnoud, and Yvonne Li.

A special thanks goes to Lulu Crisostomo, for skillfully converting administrative hurdles to (seemingly) simple tasks, and for the graceful patience with which she handles these.

On a personal note, I wish to express my sincerest gratitude to Diana Harvey, for enthusiastically participating in scientific and non-scientific discussions with me since our undergraduate years. Her willingness to consider all perspectives, remarkable recall of facts and statistics, curiosity, and ability to multi-task, have been an inspiration to my own self-improvement. I am deeply grateful to Rebecca Hunt-Newbury, whose friendship through the ups and downs of my PhD has been invaluable. I thank her for numerous insightful discussions, both in the lab and on the local hiking trails, showing me that science is not merely an activity, but a mindset.

Finally, to my family, I extend my heartfelt thanks for their patience, encouragement, and unfailing support. I thank my mother and father, Mariana and Sergiu, for encouraging my natural curiosity and consequent pursuit of a scientific career. And I thank Rick, my husband, for imbuing each day with love and laughter.

1. Bioinformatic approaches for the analysis of antisense transcript expression, evolution, and regulatory effects[1]

1.1 Introduction

Elucidating the mechanisms by which gene expression is regulated is one of the main challenges in understanding mammalian development and disease. Antisense transcription has recently been recognized as one such mechanism, and is characterized by the transcription of an overlapping gene from the opposite strand. A number of studies of individual loci have determined that gene pairs encoded in an overlapping and opposite orientation (termed sense-antisense (SAS) gene pairs) are able to regulate the processing of their partner. One way in which this regulation is exerted relies on the perfect sequence complementarity that SAS transcripts have over the region of overlap, and involves formation of double stranded RNA (dsRNA). Processing of dsRNA can alter the maturation, nuclear transport, splicing, or message stability of the SAS transcripts. Furthermore, there is substantial evidence that chromatin state, and thus sense gene transcription, may also be regulated by antisense transcripts (as described in this chapter).

To date, analysis of EST libraries and cDNA libraries, mRNA sequences, transcript tagging technologies, and microarrays have provided evidence that a large proportion (63% to over 90%) of the mouse and human genomes are transcribed into RNA (Carninci et al. 2005; Cheng et al. 2005), and further, that approximately 60% of transcripts are involved in sense-antisense pairs. Consequently, there is significant potential for widespread antisense-mediated regulation of mammalian gene expression.

This chapter reviews known examples of SAS-mediated regulation and techniques used to determine the prevalence of antisense transcription, primarily in the mouse and human genomes. Finally, I review analyses of SAS evolutionary conservation, regulated expression, and disease relevance.

[1] A version of this chapter has been published. Petrescu AS, Marra MA. 2007. *Genomic approaches for identifying cis-encoded antisense transcripts.* Global Research Network. Kerala, India. Research Advances in Nucleic Acids Research.

1.2 Thesis overview

Systematic analyses of comprehensive transcriptome datasets have revealed thousands of mammalian SAS loci. A consistent positive association between increased sampling depth and the number of detected transcripts, including antisense transcripts, indicates that only moderate-abundance and high-abundance transcripts have been comprehensively profiled (reviewed in section **1.4**). To identify all expressed sequences in the mammalian genome, it is thus necessary to devise ways of efficiently profiling low-abundance transcripts.

The first **hypothesis** of my thesis was that deeper sampling would generate a more comprehensive profiling of transcription in human cells, including antisense transcription. Comparing large collections of EST (expressed sequence tags) libraries and tag sequence libraries, which were sampled at the same depth, showed that tag sequencing performs better at identifying low-frequency expression (Chen et al. 2002). Thus, deeper sampling of tag sequence libraries is one approach by which these sequences might be efficiently discovered, however, the high cost of additional sequencing would be prohibitive. This limitation can be overcome with the application of 'next-generation' sequencing platforms, which enable ultra-high throughput sampling in a cost effective manner (section **1.4.4**). One such application (Tag-seq) was developed at the BCCA's Genome Sciences Centre, and utilized the Illumina platform to sequence tag sequence libraries generated through a modified LongSAGE protocol (**Chapter 1**). A **specific aim** of my thesis was to compare the transcript profiling success of the LongSAGE method to that of Tag-seq, which generates transcriptome libraries that are sampled an order of magnitude more deeply. Specifically, I conducted cross-platform comparisons to assess the suitability of Tag-seq for identification of low-frequency transcripts, such as antisense transcripts. A **second aim** was to use Tag-seq data to conduct an analysis of differentially expressed antisense transcripts in cancerous versus normal cells.

A regulatory role of antisense transcription involves alteration of sense gene splicing outcomes, and has been well documented at the thyroid hormone receptor locus (Hastings et al. 2000). However, to date no studies have surveyed the prevalence of antisense-mediated splicing events on a genome-wide scale. My **second hypothesis** was

that antisense transcripts may have a role in the regulation of alternative splicing at numerous loci. Consequently, a **principal aim** of my thesis was to develop a computational approach that used exon tiling array data to identify antisense-correlated splicing events genome-wide (**Chapter 3**). Exploring this relationship further, I used public nucleosome and polymerase occupancy datasets to propose a novel mechanism by which alternative splicing might influence splicing outcomes (**Chapter 3**).

Having defined a relationship between antisense transcription and alternative splicing (**Chapter 3**), and knowing that alternative expression events influence cancer biology (Venables 2004), I **hypothesized** that antisense-correlated alternative splicing events may contribute to the molecular heterogeneity of cancer types or subtypes. In **Chapter 4**, I explore the utility of using antisense-correlated splicing events to identify clinically distinct sub-types of cancer.

1.3 Biological roles of antisense transcription

Antisense transcription can affect the processing of the sense partner gene at both the transcriptional and post-transcriptional levels. Control at the transcriptional level can be mediated by either transcriptional interference or mechanisms of chromatin silencing induced by antisense expression. Post-transcriptional regulation can occur through three dsRNA-dependent mechanisms: RNA editing, RNA interference, and RNA masking (Lavorgna et al. 2004).

1.3.1 Transcriptional interference

Antisense-mediated regulation via transcriptional interference operates through premature termination or stalling of the RNA Polymerase II complex (RNAP) on one strand, due to steric interference arising from the presence of transcription machinery on the other strand (Eszterhas et al. 2002; Shearwin et al. 2005). Since transcription involves the movement of an RNAP complex along an unwinding DNA strand in the 5' to 3' direction of the gene, concurrent antisense transcription can result in supercoiling of DNA between RNAP complexes. This sterically unfavourable state can be resolved through RNAP stalling, backtracking, or disassociation (Eszterhas et al. 2002; Shearwin et al. 2005; Galburt et al. 2007).

1.3.2 Epigenetic effects

Chromatin is the structured association of DNA and nucleosomes, and its "active" (euchromatic) or "silent" (heterochromatic) state directly affects the accessibility of target DNA to protein complexes that carry out transcription. The reversal and maintenance of silent and active states partially involves the modification of specific nucleosome residues as well as DNA methylation, through mechanisms that are the topic of many current research efforts (Thiriet and Hayes 2005) for review, see (Jiang and Pugh 2009).

Chromatin silencing induced through DNA methylation and histone modifications seems to be a hallmark of imprinted genes that are controlled by antisense transcripts (Malik et al. 2000). In one study, an estimated 85% of imprinted genes were found to be associated with antisense RNA(Carninci et al. 2005)s . This suggests a role for antisense transcription in epigenetic processes. A classical example of antisense-mediated epigenetic silencing occurs during X-chromosome inactivation and involves the SAS non-coding RNAs Xist and Tsix (Lee et al. 1999). X-inactivation is a critical process during early embryogenesis in female mammals, and is carried out through Xist-mediated changes in chromatin composition. During this process, one X-chromosome is converted from active euchromatin to transcriptionally silent heterochromatin. Increased expression of Xist on the future inactive chromosome is followed by coating of the X-chromosome in Xist RNA, and an outward spread of silencing histone modifications (ex. H3K9) and CpG island DNA methylation from the Xist-proximal X inactivation center (XIC). The expression of Tsix, encoded antisense to Xist, allows one X-chromosome to escape inactivation. Transcription of Tsix through the Xist locus is required for this effect, and causes Xist inactivation through epigenetic repression of its promoter (Navarro et al. 2006).

Non-imprinted loci, such as sphingosine kinase-1 (Sphk1), can also be epigenetically regulated by antisense transcripts (Imamura et al. 2004). At this locus, tissue specific alternative splicing of Sphk1 is controlled by the methylation of a tissue-dependent differentially methylated region (T-DMR) embedded in a CpG island. An antisense transcript (Khps1a) overlapping the T-DMR can alter the methylation of the Sphk1 CpG island, and consequently influence processing of the Sphk1 gene.

Despite the documented relationship between antisense transcription and chromatin modifications, the specific events mediating antisense dependent silencing of large chromosomal regions remain to be thoroughly understood. One possibility already mentioned involves non-coding RNA-mediated targeting of chromatin modification enzymes to specific chromatin domains. This mechanism is supported by evidence linking the functional demarcation of active and silent chromatin domains in human HOX loci to a physical interaction between a chromatin-remodeling complex, the polycomb repressive complex 2 (PRC2), and an antisense transcript in the HOXC locus (HOTAIR; (Rinn et al. 2007)). HOTAIR and PRC2 interactions are necessary for PRC2 localization to the HOXD locus, and the consequent silencing of the region through trimethylation of histone 3 lysine 27 (H3K27me3) residues. Thus, in contrast to previous studies, the observed function of this antisense transcript (encoded on chromosome 12) was in *trans* (i.e. encompassing 40 Kb of DNA at the HOXD locus encoded on chromosome 2). Antisense-mediated heterochromatinization does not always occur over large distances, and may induce silencing through localized changes in the promoter of the sense gene (Yu et al. 2008).

Antisense-mediated changes involving DNA methylation have also been observed to occur in cis, and a unique example underlies a case of α-thalassemia (Tufarelli et al. 2003). The disease-causing genomic event in this patient was found to be a deletion closely juxtaposing the hemoglobin α-2 gene (HBA2) and a promoter-containing segment of a distant gene on the opposite strand, LUC7. In this patient, transcription of LUC7 carried through the opposite strand of the HBA2 gene and promoter, leading to the de novo methylation of the HBA2 promoter, and effectively silencing the gene. The exact mechanism of antisense transcription induced DNA-methylation remains to be elucidated.

Understanding the regulatory effects of antisense transcription on neighbouring genes will be critical given the extent of such transcription (Cheng et al. 2005). One recent study of growth-factor induced gene expression showed that transcription at one locus frequently affected neighbouring loci (Ebisuya et al. 2008). Specifically, intensive transcription induced by growth factors was often accompanied by coordinated upregulation of neighbouring genes. This effect was mediated by activating chromatin

marks (e.g. H3 and H4 acetylation) deposited at the target gene as well as in surrounding intergenic regions. Consequently, promoters encoded in neighbouring areas became available to the transcription machinery, and caused a coordinated expression pattern, termed a "ripple effect". This ripple effect may account for a large proportion of transcriptional events in unannotated intergenic regions, including antisense transcription. Understanding the effects of such transcription will be a complex challenge to address in future studies.

1.3.3 RNA editing

RNA editing is a process carried out by nuclear adenosine deaminases that act on RNA (ADARs). ADARs bind dsRNA, such as that resulting from long regions of SAS complementarity, and subsequently catalyze the hydrolytic deamination of adenosine residues to inosine. Because inosine is read as guanine by the transcriptional machinery, editing can alter codons, create or remove splice sites, sequester RNA molecules to the nucleus, or lead to dsRNA degradation (Bass 2002; Athanasiadis et al. 2004). Antisense-mediated RNA editing has been observed at the thymidylate synthase (TS) locus, a gene expressed in rapidly dividing cancer cells, and a major chemotherapeutic target (Johnston et al. 1994). In the presence of the antisense transcript (rTSα), the TS RNA is cleaved at specific adenosine residues and consequently down-regulated (Chu and Dolnick 2002). Cleavage at inosilated residues is likely carried out by RNases that specifically target dsRNA formed between the sense and antisense transcripts (Meegan and Marcus 1989).

1.3.4 RNAi

Gene silencing induced through dsRNA dependant activation of the RNAi pathway has been documented in both fly and yeast (Aravin et al. 2001; Volpe et al. 2002). Although no mammalian examples of RNAi induced by SAS dsRNAs have yet been described, long dsRNA introduced into mammalian cells can be processed through the RNAi pathway, and knock down expression of genes with homologous sequences (Shinagawa and Ishii 2003). RNAi is therefore a plausible consequence of SAS gene transcription. One study (Haussecker and Proudfoot 2005) presents compelling evidence for the association of antisense transcription with the RNAi pathway in humans. This study tested the hypothesis that intergenic transcription at the β-globin locus was positively associated with an open chromatin state and with globin gene expression. The results

showed that the expression of intergenic SAS transcripts was not positively correlated to either an open chromatin state, or globin gene expression. Further investigation showed that the SAS intergenic transcripts were up-regulated in cells in which Dicer was knocked down (Haussecker and Proudfoot 2005), suggesting that Dicer, a protein involved in the RNAi pathway, promotes chromatin inactivation. These observations are consistent with a model in which intergenic SAS transcription contributes to the formation and maintenance of silent chromatin, as previously observed in X-inactivation (Navarro et al. 2006), but in this instance involving the RNAi pathway. At this locus, silenced chromatin is the default state and can be reversed in the presence of transcription factors that are specifically expressed in erythrocytes, ensuring tissue-specific transcription of globin genes.

1.3.5 RNA masking

The third post-transcriptional mechanism, RNA masking, involves the formation of dsRNA encompassing regulatory regions of either the sense or antisense transcripts. The formation of dsRNA can interfere with the binding of various trans-acting proteins or RNA factors that recognize single-stranded RNA targets. Masking of sequence motifs such as splice sites, splice site enhancers and repressors, polyadenylation signals, can therefore alter RNA splicing, stability, and localization (Kuersten and Goodwin 2003). One well characterized example is the post-transcriptional regulation of thyroid hormone receptor (TRα, also known as THRA) expression by the antisense transcript Rev-erbA-α (also known as NR1D1; **Fig 1.1A** (Hastings et al. 2000)). Alternative splicing of the sense TRα pre-mRNA generates two functionally antagonistic proteins: TRα1, which mediates signals through the thyroid hormone, and TRα2, an orphan nuclear receptor that competes with TRα1 for DNA and protein binding sites. The biological response of a cell to thyroid hormone is thus dependant on the ratio of TRα1:TRα2, and Rev-erbA-α modulates this ratio by occluding the TRα2 specific splice site and skewing pre-mRNA splicing into the TRα1 form. Consequently, interactions between the sense and antisense transcripts at this locus determine the cellular response to thyroid hormone.

1.4 High-throughput discovery of SAS genes

Prior to the availability of genome sequence data, approximately 40 SAS transcripts had been identified in studies of individual loci (Kumar and Carmichael 1998; Vanhee-

Brossollet and Vaquero 1998). Our understanding of the prevalence and nature of mammalian antisense transcription has increased dramatically in the past decade, with the availability of whole genome sequences, vast collections of ESTs, cDNA libraries (complementary DNA), well-characterized mRNAs, and through the use of high-throughput transcriptome profiling techniques such as microarrays and sequence tags (**Fig. 1.2**).

1.4.1 Studies utilizing mRNAs, ESTs, and cDNA libraries

The first large-scale observations of SAS transcription were based on ESTs, cDNA, and mRNA data. cDNA libraries can be created from polyadenylated (poly(A)+) mRNAs that are reverse-transcribed into complementary DNA sequences with an oligo-d(T) primer, and subsequently cloned into vectors. The internal structure of such transcripts can be assessed by full-length sequencing of cDNA libraries. While this is costly and time-consuming, collaborative efforts have resulted in comprehensive collections of fully sequenced cDNAs (Okazaki et al. 2002; Gerhard et al. 2004; Ota et al. 2004).

Sequencing of these libraries using primers specific for the 5' and 3' flanking vector sequences rapidly and more cheaply generates EST libraries of partial clone sequences that can be aligned to a reference genome for transcript identification (Hillier et al. 1996). A limitation of EST libraries in generating quantitative measurements of transcript expression is their historically shallow depth (on the order of tens of thousands of reads) relative to the number of mRNA molecules per cell (estimated at approximately 300-500 thousand, (Jackson et al. 2000)). In addition, ESTs only provide partial 5' and 3' sequence coverage for transcripts longer than generated read lengths.

Despite these limitations, mining collections of cDNA, EST and mRNA sequence data resulted in the identification of antisense transcripts for up to 51% of genes. An initial approach to systematic SAS gene identification employed BLAST (Kent 2002) to find regions of complementarity between vertebrate mRNAs from RefSeq (Lehner et al. 2002). mRNA sequences have generally reliable orientation information, which is critical to identifying SAS transcripts, but were available in limited numbers, leading to the identification of only 87 SAS pairs, a small subset of currently known SAS pairs (**Table 1.1**). This estimate was quickly increased through the use of EST data to augment the RefSeq mRNA collection (Shendure and Church 2002). Since sequence orientation for

ESTs is sometimes nonexistent or unreliable, the search was limited to directionally cloned ESTs with splicing information and mapping near polyA signals and tails. From these data, 217 candidate cis-encoded mouse and human antisense genes were identified. In addition to the 87 pairs found by Lehner and colleagues (Lehner et al. 2002), and the 40 well-described cis-encoded antisense cases in the literature (Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998), this study brought the total number of SAS genes to more than 300.

This estimate increased by an order of magnitude with the development of the Antisensor algorithm (Yelin et al. 2003). This algorithm mined human EST and mRNA sequences from GenBank, and created 2,667 high-confidence clusters containing transcripts from one or both strands that share sequence. These clusters contained mRNAs and ESTs with an annotated direction, splice junction consensus sequences, and polyA tail sequences. Employing a similar clustering strategy, another group (Chen et al. 2004) used more than 4 million stringently filtered mRNAs and ESTs to generate 26,741 clusters of sequences that shared a genomic location and orientation. Of the total transcription clusters formed, 22% (5,880) consisted of SAS pairs.

Together, these reports highlight two fundamental aspects of the transcriptome – first that by using increasingly comprehensive transcriptome resources, increasing numbers of SAS transcripts can be identified, and second that expression from both strands of the genome is much more prevalent than was previously recognized (Vanhee-Brossollet and Vaquero 1998). Despite their success in identifying SAS pairs, these approaches were ultimately limited by a focus on protein-coding genes; a bias toward spliced transcripts of annotated genes; and were inherently limited to finding those SAS transcripts overlapping in exonic regions.

To overcome these limitations, the search for antisense transcripts was expanded to include non-coding RNAs, SAS pairs overlapping on introns as well as exons, and to include un-spliced transcripts. As described in the previous sections, antisense transcripts have the potential to affect sense transcription even in an unprocessed state in the nucleus; for instance via transcriptional interference or through epigenetic modifications. Thus, non-coding antisense transcripts or those overlapping sense introns are legitimate candidate targets for identification, as they may play regulatory roles. The prevalence of

antisense transcripts differing in coding status and the presence of intronic overlaps was addressed in an analysis of the Fantom2 full-length cDNA collection, in which non-coding RNAs are well represented (Kiyosawa et al. 2003). Together with mRNA data and the mouse reference genome sequence this resource allowed identification of 2,481 exon-overlapping SAS pairs and 899 intron-overlapping SAS pairs. Interestingly, approximately a quarter of the SAS transcript pairs consisted of two coding transcripts, half consisted of one coding and one non-coding transcript, and the remainder consisted of two non-coding transcripts. This evidence suggests that up to 50% of antisense transcripts are non-coding RNAs. The difference in the magnitude of SAS predictions between this and previous studies again highlights that with increased expressed sequence data, the estimation of SAS pairs grows.

1.4.2 Microarray studies

Microarrays are a well-established method for simultaneously detecting the level of expression of thousands of transcripts in a sample. Briefly, they consist of a surface (i.e. a "chip") arrayed with short probes (25-60 nucleotides) that are complementary to non-repetitive target regions of interest. Commercially available arrays can have 10-20 probesets per gene, but detection of expression is generally biased to the 3' end (**Fig. 1.2**; reviewed in (Schulze and Downward 2001)). More recent design strategies involve the tiling of probesets at regular intervals (5-35bp) over large, non-repetitive genomic regions. The advantage of tiling arrays is the ability to detect *de novo* transcription throughout the genome. However, due to the large size of the human genome, initial tiling arrays only profiled chromosomes 21, and 22 (Kapranov et al. 2002; Rinn et al. 2003; Kampa et al. 2004). Subsequent tiling array designs covered 10 chromosomes at 5bp resolution (Cheng et al. 2005), and the whole genome at 50bp resolution (Bertone et al. 2004). In addition to known annotations, these arrays are capable of detecting novel genes, novel exons, and un-annotated alternative exon boundaries. However, the vast number of chips required to tile the genome (52 million probesets spanning 134 chips) is a prohibitive limitation in applying this approach to routine gene expression measurements. The balance between comprehensive profiling and frugal use of space was therefore a design feature of exon tiling arrays, which have 25bp probes spanning 1.4 million known and predicted exons in the human genome (www.affymetrix.com). These fit on one chip, allowing high-throughput and low-cost quantitative experiments.

Since these arrays are not designed to elucidate exon connectivity of alternate splice variants at a given locus, additional methods have been developed to create arrays capable of detecting known and predicted exon splicing events (Nuwaysir et al. 2002; Johnson et al. 2003; Griffith et al. 2010).

A major finding of the tiling arrays profiling human chromosomes 21 and 22 revealed that a surprisingly large proportion of the intragenic and intergenic regions are in fact transcribed (Rinn et al. 2003; Kampa et al. 2004), and that intergenic antisense transcripts and novel exons in known genes were expressed in approximately equal proportions (Bertone et al. 2004). This was in accordance with the findings of Rinn and colleagues (Rinn et al. 2003), who showed that half of the chromosome 22 transcribed regions overlapping introns were in an antisense orientation, and half were likely novel exons. It is possible that these antisense transcripts eluded prior detection due to their relatively low abundance, or due to increased sensitivity relative to previous methods.

Previous sequence-based estimates of SAS prevalence (>50% of genes) were confirmed in a tiling microarray experiment surveying ten human chromosomes (Cheng et al. 2005). In this analysis, ~50% of the detected transcribed sequences did not overlap with any of the well-characterized exon, mRNA, or EST annotations. Because this microarray design did not allow the strand of origin to be distinguished for a given transcribed sequence, a combination of RACE, high-density arrays, and cloning and sequencing techniques was used to characterize a subset of 768 transcripts in detail. The majority (60.8%) of surveyed loci for which 5' and 3' RACE was successful had evidence of transcription on both strands (Kapranov et al. 2005).

1.4.2.1 PolyA(-) antisense transcripts

Non-polyadenylated (poly(A)-) transcribed sequences constitute nearly half of the transcriptome (Cheng et al. 2005), yet the vast majority of sequence resources and high-throughput array experiments only focus on poly(A)+ mRNAs. In one study addressing both poly(A)+ and poly(A)- transcripts (Cheng et al. 2005), RNA isolated from the nuclear and cytosolic compartments of HepG2 cell lines was profiled on tiling arrays spanning 10 chromosomes (~30% of the genome). Previously considered "junk" genomic regions were found to encode multiple overlapping poly(A)+ and poly(A)- coding and non-coding transcripts. Novel poly(A)- transcripts actually comprised the

major proportion of the transcribed messages in the human genome. Of all transcribed sequences, 19.4% were observed to be poly(A)+, 43.7% were poly(A)-, and 36.9%, were bimorphic. Bimorphic transcripts are transcribed as poly(A)+ RNAs and later processed to reduce or completely remove the 3' polyA tail. The specific conditions and signals regulating the bimorphic state may thus be relevant to the regulation of antisense transcripts. It remains to be determined what proportion of transcribed loci consists of poly(A)- antisense transcripts. Because the array detected transcription in a strand-independent manner, the full-length structures of many transcriptional units were determined using RACE and high-density arrays, but only for poly(A)+ RNAs. Thus, the potential roles played by poly(A)- RNAs in cellular processes may be underappreciated, and because many antisense transcripts are found to be poly(A)-, a comprehensive understanding of SAS gene regulation will ultimately require the study of this subset of the transcriptome.

A different approach was taken by Kiyosawa and colleagues (Kiyosawa et al. 2005), who used custom microarrays to investigate the expression of 1,947 known SAS pairs. To determine the proportion of poly(A)- RNAs transcribed from these loci, the expression values of RNA primed with random nanomers were compared to those of RNA primed with poly(T) primers. The results obtained with random nanomers were expected to differ from the poly(T) results only if the majority of SAS pairs were not polyadenylated. Interestingly, random-primed targets gave higher signals, an effect specific to SAS genes but not non-overlapping genes. The poly(A)- fraction of the transcriptome may therefore harbour regulatory SAS RNA species unlikely to have been previously detected in EST and cDNA libraries. The existence of poly(A)- SAS transcripts is consistent with regulatory roles carried out in the nucleus, involving for instance antisense-mediated altered splicing of sense pre-mRNA.

1.4.3 Tag-based studies

Serial Analysis of Gene Expression (SAGE) (Velculescu et al. 1995), is a technique used for gene expression profiling and *de novo* transcript discovery, without need for probe design or prior knowledge of either genomic sequence or transcribed regions, as required for microarray design. A major aim of developing this technique was to increase the efficiency of sequence-driven transcript profiling by increasing the number of transcripts

detected per sequence read. To generate SAGE libraries, a biotinylated poly(T) primer is used to synthesize cDNA from an mRNA sample. The cDNAs are tethered to streptavidin-coated beads and digested using an enzyme that cleaves typical transcripts at least once (ie. NlaIII). Tethered cDNA fragments are ligated to 5' adapter sequences that contain a recognition sequence for a type II restriction enzyme. This tagging enzyme cuts several nucleotides away, generating short tags that can be cloned into vectors, amplified using PCR, and subsequently sequenced. The SAGE technique thus results in the generation of sequence reads containing 30-45 short sequence tags (depending on read length) from the 3' ends of poly(A)+ transcripts, which can be mapped to genome and transcript resources to provide digital counts of transcript expression (Velculescu et al. 1995). To assess the expression of rare transcripts that are poorly represented in EST libraries is a simple matter of increasing the number of sequences analyzed.

Different type II restriction enzymes have been used to generate different length tags of either 14bp (BsmFI in short SAGE, (Velculescu et al. 1995)), or 21bp (MmeI in LongSAGE,(Saha et al. 2002)). However, due to the likelihood of finding a random 14bp sequence multiple times in the genome, tags generated using the short SAGE technique can be mapped to unique locations on the genome in only a small fraction of cases, and are instead most frequently mapped only to transcript sequences. The LongSAGE technique was developed to address this limitation, and generates 21bp tags with a 75% unique mapping rate to the genome (Saha et al. 2002). Collections of classical (short) SAGE and LongSAGE libraries are available as part of the Cancer Genome Anatomy Project (CGAP) (Lal et al. 1999; Khattra et al. 2007).

The search for antisense transcripts using short SAGE data was pioneered by Quere and colleagues (Quere et al. 2004), who investigated whether unmapped tags in the human leukemia cell line U937 could instead be mapped onto the reverse complement of well-annotated mRNAs. A total of 3.5% of the 4,444 mRNAs with detectable expression in U937 cells showed expression of both sense and antisense tags, and 2.8% showed expression of the antisense tag only. By definition, the antisense tags that mapped to a sense gene sequence represented a convergent antisense gene (ie. whose 3' end overlapped the sense gene 3' end; **Fig. 1.2**). Therefore, the 6.3% of genes with antisense transcription were an underestimate of the total antisense transcription in these cells, as

SAGE tags from pairs with intron overlaps and from divergent SAS pairs were not considered.

The first study to undertake discovery of antisense transcription in mouse embryonic tail tissues using LongSAGE employed a stringent tag-to-gene mapping process and required that all tags be supported by EST or cDNA evidence (Wahl et al. 2005). Antisense tags mapping outside of sense gene boundaries (such as those arising from divergent genes, **Fig. 1.1**) were included if they mapped within 10kb of gene boundaries and had supporting EST or cDNA evidence. Of the 1,260 genes with detectable antisense transcription, only 259 had annotated antisense Ensembl transcripts. Thus, >75% of the antisense transcription detected using LongSAGE in one murine tissue was novel, indicating that the number of antisense transcripts detected should dramatically increase in larger LongSAGE libraries generated from multiple tissues.

Such an analysis was carried out using 8.55 million LongSAGE tags derived from 72 LongSAGE libraries as part of the Mouse Atlas of Gene Expression project (Siddiqui et al. 2005). Libraries were sampled to an average depth of >118,000 tags, yielding gene-detection sensitivity approximately equivalent to that of microarray approaches (Su et al. 2002). Thus these libraries were suitable for detection of abundant and moderately abundant transcripts. After removing tags with sequence errors and applying a tag quality threshold, 261,134 tag sequences were mapped uniquely to the genome. Of these, 46% (120,122) mapped to 19,865 known genes, and 20% (52,255) mapped antisense to annotated genes. Thus, over a third of mapped sequence tags appeared to derive from antisense transcripts.

A different tagging technology, Cap Analysis of Gene Expression (CAGE), can be used to generate digital tag counts of capped mRNAs (Kodzius et al. 2006). This technique is similar to LongSAGE, but the ~20bp tags are generated from the 5' ends of transcripts and can be prepared from total RNA, allowing identification of both poly(A)+ and poly(A)- transcripts. The 5' cap of a mature transcript is a modified guanosine nucleotide that prevents transcript degradation by exonucleases and regulates nuclear export of mRNAs (Wilusz et al. 2001)**.** Essentially, CAGE libraries profile transcriptional start sites of mRNAs in a sample, and provide a quantitative digital measure of mRNA expression. Together, SAGE and CAGE tags allow comprehensive but independent

detection of transcription start sites as well as terminus-proximal sites (i.e. 3'-most NlaIII sites). These sites can be simultaneously profiled using a third tagging method, GIS (gene identification signature) (Ng et al. 2005). This approach generates linked 5' and 3' paired end tags (PET) from one mRNA molecule, allowing analysis of transcript isoforms differing in both start and end exon usage. Generation of PET tags involves the circularization of cDNAs into cloning vectors, which can introduce a size-bias in clones. Consequently, PET tags are not generally used for quantitative analyses; instead, they can be used to annotate transcript termini in a library that is profiled in parallel with a quantitative technology such as CAGE (Carninci et al. 2005).

The RIKEN group compiled an exhaustive profile of the mouse transcriptome using sequence-based expression data including cDNA and EST libraries, GIS and CAGE tags, and found that depending on stringency, between 28.7% and 72.1% of mouse transcripts had evidence for expression from the opposing strand (Riken Genome Exploration Research et al. 2005). In the most stringent case, cDNAs provided evidence for antisense transcription at 28.7% of loci. For 51.2% of expressed loci, antisense transcription was detected by at least two independent observations of expression (e.g. ESTs). Requiring just one observation for expression increased this number to 72.1%. These findings support the previous observations that antisense transcription is pervasive in the mammalian genome, and that a significant proportion of that transcription is near the limit of detection using these methods.

1.4.4 Profiling antisense transcripts using next-generation approaches

As noted in the previous section, antisense transcripts can often be infrequently expressed, requiring increasingly deeper sampling for the detection of novel low-abundance sequences. Additional insight into the prevalence of antisense transcripts can therefore be gained by the application of methods capable of sensitively sampling low-abundance transcripts. The development of 'next generation' sequencing platforms has made possible a cost-effective increase in the throughput of sampling by at least an order of magnitude relative to Sanger-based methods. These ultra-high throughput methods include the Roche Applied Sciences' 454 pyrosequencing platform (Margulies et al. 2005), the Illumina Genome Analyzer (Bentley 2006), and Applied Biosystems SOLiD platform (Valouev et al. 2008), and rely on massively parallel production of short reads.

Each of these technologies generates libraries of fragmented input sequences annealed to platform specific adapters, and uses the adapters to perform PCR amplification of single stranded sequences. In the pyrosequencing approach, subsequent DNA sequencing involves the release of a pyrophosphate upon nucleotide incorporation, which ultimately leads to detectable light emission. Individual fluorescently-labeled bases are sequentially washed over the single stranded templates, and incorporation events are detected for millions of templates in parallel. The intensity of light produced is proportional to the number of incorporated nucleotides. Thus, one limitation of this approach is light-detector saturation during a homopolymer run, which leads to errors in the estimated number of identical bases. In this platform, single-stranded templates are each bound to individual agarose beads, and undergo the PCR and sequencing reactions in isolation.

The Illumina platform performs the parallel PCR amplification of millions of adapter-ligated single-stranded sequences on a single flowcell. This process leads to the generation of millions of clusters, each containing a million PCR-produced copies of a distinct single molecule fragment. This redundancy allows efficient imaging of nucleotide incorporation events. Each type of nucleotide has a distinct fluorescent label, allowing their simultaneous detection. Sequencing-by-synthesis proceeds from flowcell-attached primers complementary to the library adapters. A chemical block prevents more than one base from being added to the complementary DNA in each incorporation reaction. The subsequent imaging step is thus followed by a de-blocking step.

The SOLiD platform performs emulsion PCR amplification of DNA fragments individually attached to magnetic beads. The beads are distributed on a slide surface, where sequencing proceeds by ligation of fluorescently-labeled probes, such that each sequenced base is detected twice by independent probes.

In contrast to the multi-step process of sequencing clone-based libraries using the Sanger method, these platforms allow simpler processing of sampled mRNAs into sequence reads without a bacterial cloning step. In general the reads generated by the next generation platforms are considerably shorter than those generated by capillary sequencing (36-500bp, depending on the platform, versus the ~1kb achievable by capillary sequencing). However, the shorter read lengths are amenable to tag sequencing applications, which can rely on the massive production of reads representing individual

tags instead of on long sequence reads of clone-based tag concatamers. Modification of the GIS method to utilize the Roche 454 pyrosequencing platform has demonstrated a 100-fold increase in sampling efficiency (Ng et al. 2006), showing the utility of next generation methods in improving transcript detection.

One important limitation of the data generated by these technologies to date is the lack of strand specificity. Libraries generated using the Illumina and SOLiD platforms are generally constructed in non-strand specific manner, and the strand of origin of the resulting sequences can not be easily determined. This limitation can be overcome through modified library generation methods (discussed in section 5.1), or through technologies that allow sequencing of single molecules, such as the Helicos system (Bowers et al. 2009). The Helicos method uses reversible terminators with tethered inhibitors, and allows sequencing by synthesis of single molecules without amplification. During the course of this thesis, the study of sense-antisense transcripts using such strand-specific next generation technologies was not practical due to insufficient numbers of existing libraries.

1.5 Functional analyses

One potential caveat of SAS detection methods that rely on reverse transcription is the potential for artifactual antisense sequences generated during library preparation (Johnson et al. 2005). Thus, in addition to genomic prevalence, other lines of evidence are required to ascertain the biological relevance of antisense transcripts. Ample evidence to support the putative significance of antisense transcription has consequently been collected through studies of evolutionary conservation, association with gene regulatory motifs, and surveys of regulated expression across biological samples or in response to stimuli.

1.5.1 Evolutionary conservation

If two neighboring genes share a SAS relationship and that topological arrangement is required for regulation of gene processing, then events disrupting their topology should be selected against. Rearrangements and genome expansion events can disrupt SAS topology and their putative regulatory interactions, and should therefore be selected against to a greater extent than rearrangements at non-overlapping genes, which do not

have regulatory interactions based on sequence overlap. To test whether this selection could be observed, Dahary and colleagues (Dahary et al. 2005) compared the topological relationships of SAS genes in the human genome to those in the Fugu genome. The human genome is eight-fold larger than the Fugu genome, an expansion that can be largely accounted for by transposable elements (TEs). TEs were active at stages prior to and during the radiation of mammals, and now constitute at least 45% of the human genome (Lander et al. 2001). The insertion and movement of TEs in the mammalian lineage has presumably affected both intergenic distances and gene order. To investigate the relationship between antisense transcription and the intergenic distance of consecutive genes in the human genome, 453 gene pairs with conserved linkage between human and Fugu were identified. For this entire set, the average distance between paired genes was 11-fold larger in the human genome relative to Fugu, which is consistent with the difference in genome sizes. Interestingly, gene pairs on the same strand (170 pairs) had a 13-fold larger distance in human versus Fugu, while SAS genes (283 pairs) had a 2.5-fold larger distance. Therefore, the reduced distance observed at SAS genes indicates that selection seems to be acting to preserve the topological relationships of SAS genes versus genes encoded in tandem.

To further examine the effect of antisense transcription on the preservation of gene order, the same group used the Antisensor algorithm to determine that 236 of 2,737 consecutive gene pairs in the human genome formed SAS pairs (*52*). Antisense overlaps could not be detected in Fugu due to limited numbers of full-length cDNAs and ESTs available at the time. Of the human SAS pairs, 23.3% remained consecutive and preserved their orientation in Fugu, and were thus likely to form SAS pairs in Fugu. Conversely, only 13.5% of the same-strand human gene pairs remained consecutive in Fugu, showing that SAS genes preserved their order significantly more often than neighboring non-overlapping genes.

A substantial fraction of SAS genes overlapping on exons was found to be conserved between the mouse and human genomes (Engstrom et al. 2006). A total of 16% (962) of human SAS pairs were found to retain their overlap patterns in mouse, while 18% (943) of mouse SAS pairs retained their overlap patterns in human. To account for the variable effect of sequence resource comprehensiveness on SAS pair identification, the entire

human dataset was compared with random samples of varying sizes of all available mouse transcript sequences. The resulting saturation curve predicted that approximately 25% of human SAS pairs were conserved in mouse. The analogous analysis, sampling human transcripts instead of mouse transcripts, generated a saturation model predicting that approximately 26% of mouse SAS pairs were conserved in human. Thus, the exonic overlap patterns of a quarter of all exon-overlapping SAS gene pairs were evolutionarily conserved (Engstrom et al. 2006; Zhang et al. 2006). This implies a common functional role for at least a subset of antisense transcripts, which constrains their separation.

Although up to 75% of the human and mouse SAS gene overlaps are organism-specific, a lack of conservation at the sequence level does not imply a lack of function. As described, a significant proportion of antisense transcripts are non-coding RNAs that have sequence-independent regulatory functions. It is thus likely that antisense-strand transcription itself has been the target of positive selection.

1.5.2 Regulated expression

Non-coding genes by definition do not have a primary function as proteins, thus they are likely to be involved in transcriptional and post-transcriptional regulation (Mattick 2004). Since approximately 50% of SAS genes are non-coding (Cawley et al. 2004; Riken Genome Exploration Research et al. 2005), assessing their functional attributes is a biologically relevant challenge. Two functional attributes amenable to testing are the presence of transcription factor binding sites (TFBS) in regulatory regions, and differential non-coding RNA expression in response to stimuli.

An analysis of these functional attributes was conducted for coding and non-coding transcripts on chromosomes 21 and 22 (Cawley et al. 2004). First, chromatin bound by the transcription factors (TFs) Sp1, cMyc, and p53, was isolated using chromatin immunoprecipitation (ChIP), and analyzed on tiling arrays (ChIP-Chip). Unexpectedly, only a minority (22%) of TFBSs were located within canonical promoter regions in the 5' UTRs of coding genes. A larger proportion (36%) were found within or just 3' to well-characterized genes, indicating they might have a role as distal regulatory elements active on protein coding genes (e.g. enhancers), or as promoters for non-coding transcripts such as antisense transcripts.

Overall, similar proportions of coding and non-coding genes were associated with TFBSs. Evidence specifically pertaining to 363 SAS loci revealed a strong association of antisense transcription with non-canonical TFBSs, half of which were evolutionarily conserved between mouse and human.

Next, the propensity of non-coding transcripts to respond to differentiation signals was investigated by challenging cells with retinoic acid (RA). Non-coding antisense transcripts exhibited a strong pattern of differential expression that was highly concordant to the RA response of corresponding coding sense genes, with which they were highly co-expressed.

In keeping with these observations, multiple studies confirm that that co-expressed SAS pairs occur at a greater than expected frequency (Chen et al. 2004; Riken Genome Exploration Research et al. 2005), and that the patterns of expression can be tissue specific (Kiyosawa et al. 2005), implying that non-coding antisense RNAs may be involved in tissue-specific regulation of sense gene expression.

Direct testing of SAS interactions can be achieved by targeting siRNAs against non-overlapping regions of each partner in a SAS pair (Riken Genome Exploration Research et al. 2005). Results from such perturbation studies underscore the underappreciated complexity of SAS regulation. In one case, siRNA inhibition of the antisense transcript led to an increase in sense mRNA, but siRNA inhibition of the sense transcript had no effect on the antisense mRNA. In another case, inhibition of the sense transcript led to decreased levels of the antisense mRNA, however, inhibition of the antisense transcript had no effect on the sense mRNA. Surprisingly, over-expression of the sense mRNA in this case induced expression of the antisense mRNA. The interactions between SAS transcripts are thus more complex than a simple inverse expression or co-expression model would suggest, and interrogations of these interactions will require orthologous strategies.

1.6 Cancer

Abnormal and un-controlled cell division leads to cancer. The cancer phenotype is an indicator of aberrations in six aspects of cellular physiology, specifically: growth-signal independent proliferation, autonomy from anti-growth signals and programmed cell

death (apoptosis), sustained angiogenesis, limitless replicative potential, and the ability to invade other tissues, giving rise to metastases (Hanahan and Weinberg 2000).

In British Columbia, the cancer incidence rate is 774 per 100,000 individuals, with an associated mortality rate of 305.5 per 100,000 individuals. In 2006 over 17,000 cases of cancer were diagnosed in the province (www.bccancer.bc.ca), underscoring the significant burden that this disease places on the medical system, and the importance of this area of health research.

1.6.1 SAS transcripts in human disease and cancer

Numerous studies have reported antisense transcripts to disease-related genes, and significant changes in the ratio of sense to antisense transcripts in disease tissue (Krystal et al. 1990; Smilinich et al. 1999; Thrash-Bingham and Tartof 1999; Chu and Dolnick 2002; Rossignol et al. 2002; Shendure and Church 2002; Mihalich et al. 2003; Reis et al. 2004; Shirasawa et al. 2004; Alfano et al. 2005; Chen et al. 2005; Yan et al. 2005; Ladd et al. 2007; Yu et al. 2008). Understanding the regulatory role of antisense transcripts in regulating known cancer genes ((Yu et al. 2008), and references therein) is of particular interest. Insights into such regulatory mechanisms could provide a new understanding of relevant cancer-related regulatory mechanisms, may supply potential new targets for therapy, or serve as prognostic or therapeutic markers. For instance, expression of the antisense gene Saf has been shown to correlate with splicing of the sense partner, Fas. Fas-induced apoptosis is a key mechanism of homeostasis in the central and peripheral immune systems, where deregulation of programmed cell death can lead to autoimmunity and cancer. Although the mechanism is not completely understood, Saf inhibits Fas-mediated apoptosis. Interestingly, over-expression of Saf is correlated to specific Fas isoforms that exclude either the trans-membrane domain or the death domain, or both (Yan et al. 2005). Saf-correlated splicing events that generate a soluble rather than membrane-bound form or that skip the death domain are therefore linked to loss of function in this protein.

1.6.2 Defining cancer subtypes using microarrays

Profiling of cancer samples by microarrays generates individual measurements of thousands of gene expression values that can be used to group (i.e. cluster) patients into

groups defined by similar expression profiles. In **Chapter 4**, I use a subset of expression profiles to cluster patients into distinct groups. Numerous methods have been devised for this purpose (reviewed in (Butte 2002)), but a commonly used strategy is to first assess similarity between either genes or samples using the Pearson coefficient, and then to generate groups of similarly expressed genes using hierarchical clustering. Hierarchical clusters can be constructed by iteratively grouping the most similar genes together into larger groups, which are then grouped in yet larger groups along with other similar groups, until a complete similarity-based relationship is generated for all samples (or genes). The output of this method is based on existing information in the data, and is termed "unsupervised". In contrast, "supervised" methods are closely related but fundamentally different in that criteria for groups are first defined (i.e. cancer vs non-cancer; good pronosis vs poor prognosis), and expression data is subsequently mined to find genes that significantly differ in expression between pre-defined groups. Significance of expression differences is generally based on four characteristics: the absolute expression of the gene, the difference in expression levels between samples, the ratio of expression between groups, and the reproducibility of measurement (i.e. whether expression values are similar in a given group). Methods typically used for classification are reviewed in (Butte 2002). Successful identification of both known and novel cancer subtypes has been achieved using both unsupervised and supervised methods (Golub et al. 1999; Lubitz et al. 2006; Phillips et al. 2006; Li et al. 2009; Verhaak et al. 2010), and highlights the power of these approaches in improving diagnostic precision.

1.7 Thesis objectives and chapter summaries

The broad aims of this thesis were to (1) determine whether next generation sequencing platforms could be used to improve the quantitative detection of antisense transcripts, and (2) to explore the putative role of SAS transcription in splicing regulation and human disease.

Systematic analyses of extensive transcriptome datasets (including EST libraries, microarrays, and tag sequence libraries) have revealed that a large proportion of antisense transcripts are infrequently expressed. A principal aim of the thesis was to determine the prevalence of antisense transcription using more sensitive methods for gene expression. I took advantage of the recent development of Tag-seq to address this

aim. Tag-seq is a version of the LongSAGE protocol modified to take advantage of the ultra high sequencing throughput of the Illumina platform, and was developed at the GSC in order to complete data collection for the Cancer Genome Anatomy Project (CGAP; cgap.nci.nih.gov). In **Chapter 2**, I compared multiple Tag-seq to LongSAGE libraries, and found an improved representation of low abundance sequences, such as antisense transcripts, transcription factors, and novel exons. The majority of low-abundance transcripts found were below the levels of detection achievable by the largest LongSAGE libraries, indicating that Tag-seq significantly advances our ability to measure transcription using tag sequencing. Tag-seq detected similar numbers of transcripts to RNA-seq, but in contrast retained strand of origin information, allowing non-ambiguous identification of overlapping sense and antisense transcripts. In addition, Tag-seq outperformed microarrays in terms of detectable dynamic range, and had less GC-bias than the LongSAGE method. The digital expression counts generated through the Tag-seq approach were amenable to differential expression analysis between cancer and normal CGAP libraries, and led to the identification of SAS pairs with large changes in sense to antisense expression ratios between normal and diseased states.

Most genomic loci are now known to generate multiple transcript isoforms differing in exon usage, polyadenylation status, and transcriptional start and end sites. Alternative splicing is thought to play a role in the generation of transcript isoforms from as many as 75% of human loci, including SAS loci. Resulting transcript isoforms can encode proteins with altered localization, stability, or biological function, effectively expanding the information content of the genome into a surprisingly diverse proteome. Antisense transcripts have been shown to affect sense gene splicing outcomes (section **1.3.5**), and thus, a specific goal of my thesis was to explore the relationship between antisense transcription and alternative splicing on a global scale. In **Chapter 3,** I describe a bioinformatic approach to this challenge, based on Affymetrix exon array data. These data are both strand specific, and unlike sequence tag-based methods, provide exon-level expression information. Using 176 exon arrays profiling normal human tissues, I detected a widespread and consistent relationship between alternative splicing and antisense transcription at the majority (~75%-80%) of expressed known and novel SAS genes. To better understand the basis for this relationship, I further explored the properties of SAS sequence overlaps, regions in which alternative splicing events seemed

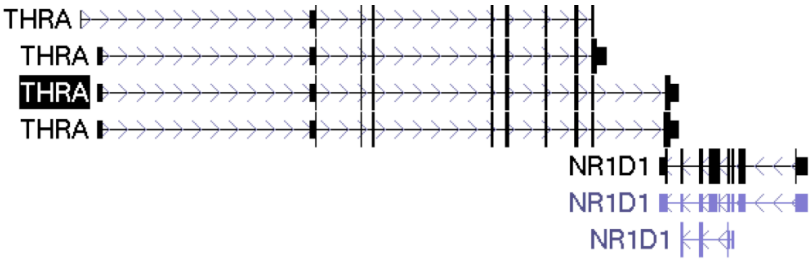
to be more prevalent. SAS sequence overlaps were distinguished from flanking non-overlapping genic regions by a significantly increased frequency of exons. I utilized nucleosome and polymerase occupancy datasets to show that a likely consequence of high exon frequencies was a concomitant increase in nucleosome occupancy levels, and a decrease in polymerase speed. Since slower polymerase elongation rates have been associated with increased alternative splicing rates (references in **Chapter 3**), my findings suggest for the first time that antisense-mediated splicing may happen through mechanisms other than RNA masking.

In **Chapter 4**, I explored the utility of antisense-correlated sense gene isoforms in understanding disease heterogeneity. Using over 1,000 exon arrays profiling normal and cancerous tissues, I first identified thousands of exons with antisense-correlated splicing events specific to cancer. I next used unsupervised hierarchical clustering methods to show that the splicing patterns of these exons could be used to cluster patients into clinically distinct groups. Specifically, groups of patients with good and poor prognosis could be identified using this approach, as well as subsets of patients that were either sensitive or resistant to standard chemotherapy. These results show for the first time that the subset of splicing events that are correlated to antisense transcription can be used to discern biologically relevant subtypes of cancer.

Figure 1.1 Regulatory coding and non-coding antisense transcripts.

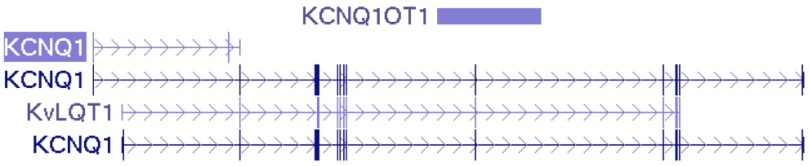
UCSC gene models for five types of SAS gene pairs with known antisense regulatory functions. These gene pairs differ in their overlap type (convergent, divergent, fully overlapping; gene direction denoted by arrows), and coding status (thin exon blocks denote non-coding regions, thicker exon blocks denote coding regions). (A) Convergent, coding / coding. (B) Convergent, unspliced non-coding / spliced non-coding. (C) Fully overlapping, unspliced non-coding / coding. (D) Divergent, coding / spliced non-coding. (E) Fully overlapping, spliced non-coding / coding, with no exon overlaps. UCSC gene models are based on RefSeq, UniProt, GenBank, CCDS, and Comparative Genomics; gene colors denote level of experimental support (http://genome.ucsc.edu/).

A



B

C


KCNQ1OT1
KCNQ1
KCNQ1
KvLQT1
KCNQ1

D


HAS2
HAS2AS

E

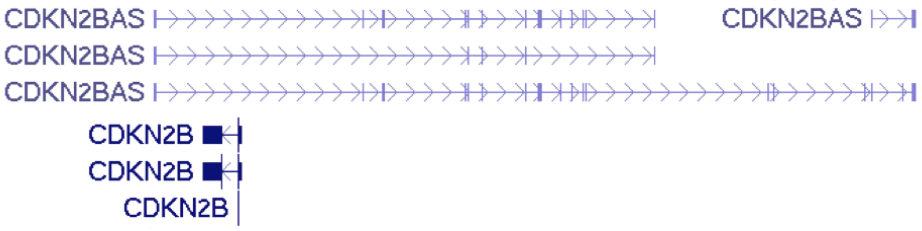

CDKN2BAS
CDKN2BAS
CDKN2BAS
CDKN2BAS
CDKN2B
CDKN2B
CDKN2B

Figure 1.2 Experimental methods for the detection of SAS transcription.

A divergent sense and antisense gene pair is depicted on the positive and negative strands of the genome. Exons (boxes) are linked by thin lines (introns) with arrows denoting orientation. Strand specific methods are displayed separately above the sense and below the antisense gene models (cDNAs and ESTs, Exon tilling arrays, SAGE and CAGE tags, and GIS tags). Methods that are not strand specific are summarized in the middle (unspliced ESTs, whole genome tilling arrays, and first generation expression arrays). Abbreviations: cDNA (complementary DNA); EST (expressed sequence tag); SAGE (serial analysis of gene expression); CAGE (cap analysis of gene expression); GIS (gene identification signature). SAS overlap is denoted by a vertical grey bar.

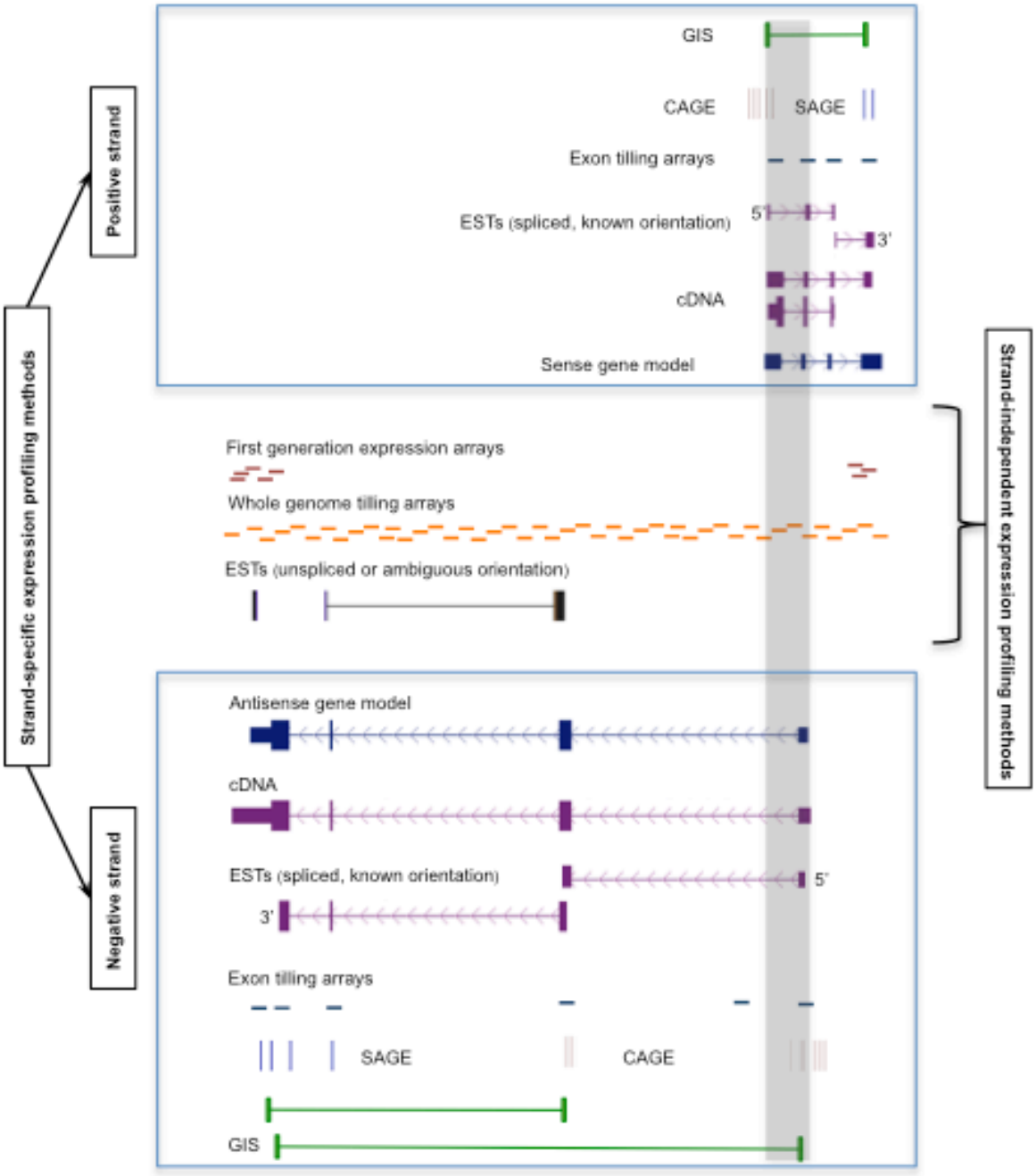


Table 1.1 EST, mRNA and cDNA based SAS analyses.

A compilation of studies relying on different types of transcript profiling methods to identify SAS genes that overlap on exons (eSAS) or introns (iSAS). Coding status of genes is summarized as c/c (both genes are coding, nc/nc (both genes are non-coding), and c/nc (one gene is coding). Topology of genes can be divergent (D), convergent (C), and fully overlapping (F). (A) Studies relying on EST, mRNA, and cDNA libraries; (B) studies relying on microarray datasets; (C) studies relying on sequence tags.

A

Dataset	Organism	Antisense transcription	Coding status	Topology	Reference
RefSeq and Ensembl mRNAs	Human	2% (87 pairs) are SAS in 12,897 RefSeq mRNAs analyzed	44% of transcripts are SAS with exon overlaps	NA	(Lehner et al. 2002)
899 human and 176 mouse EST libraries, UniGene mRNAs	Human Mouse	143 human and 73 mouse SAS pairs	NA	NA	(Shendure and Church 2002)
GenBank mRNAs and ESTs	Human	2,667 SAS pairs	NA	D: 31% C: 41% F: 28%	(Yelin et al. 2003)
37,086 TUs from Fantom2 cDNA clones; public mRNAs	Mouse	2,481 are eSAS 899 are iSAS	c/c: 27% eSAS c/nc: 55% eSAS nc/nc: 18% eSAS	NA	(Kiyosawa et al. 2003)
43,553 TUs from Fantom3; GenBank cDNAs, ESTs, CAGE, GIS	Mouse	Strong evidence: 29% Good evidence: 51% Weak evidence: 72%	c/c: 37% eSAS, 26% iSAS c/nc: 55% eSAS, 60% iSAS nc/nc: 8% eSAS, 14% iSAS	D: 36% C: 34% F: 30%	(Riken Genome Exploration Research et al. 2005)
26,741 clusters of 386,415 mRNAs and ESTs from UniGene	Human	22% eSAS and 11% iSAS pairs	c/c: 60% eSAS, 45% iSAS c/nc: 30% eSAS, 45% iSAS nc/nc: 10% eSAS, 10% iSAS	NA	(Chen et al. 2004)
36,606 TUs from ESTs, Fantom3 cDNA clones, mRNAs	Human: eSAS 24.7%, iSAS 22.7% Mouse: eSAS 25.3%, iSAS 16.1%		NA	D: 29% (Human), 33% (Mouse) C: 29% (Human), 33% (Mouse) F: 42% (Human), 33% (Mouse)	(Engstrom et al. 2006)
mRNAs and ESTs from UniGene	Human: eSAS 26.3% Mouse: 21.0% Fly: 16.8% Nematode: 2.8% Sea squirt: 15.8% Chicken: 6.6% Rat: 4.5% Frog: 4.3% Zebrafish: 2.2% Cow: 3.8%		NA	Convergent SA pairs predominate in fly, nematode, and sea squirt	(Zhang et al. 2006)
mRNAs and ESTs from UniGene	Human: eSAS (all) 22.7%, eSAS* (20,000) 2.1% Mouse: 11.6%, 2.9% Rat: 4.8%, 2.7% Chicken: 4.8%, 4.2% Fly: 17.2%, 12.1% Nematode: 0.5%, 0.5%		NA	NA	(Sun et al. 2006)

* In this analysis, the proportion of SAS genes was calculated for all expressed transcripts, or a subset of 20,000 transcripts

B

Dataset	Organism	Antisense transcription	Coding status	Topology	Reference
High-density oligoarrays of chromosomes 21 and 22. Binding sites of Sp1, c-Myc, and p53	Human	21% of transcript clusters with mRNA evidence and a non-canonical TFBS are SAS pairs	NA	NA	(Cawley et al. 2004)
13,889 transcriptional units. High density oligoarrays of the non-repetitive regions of the genome	Human	1,529 (11%) transcripts were antisense to introns	NA	NA	(Bertone et al. 2004)
19,525 transcriptional units. PCR-oligoarray of chromosome 22	Human	518 (9.8%) of 5,264 transcribed fragments were antisense to introns.	NA	NA	(Rinn et al. 2003)
Cytosolic and nuclear poly(A)+/- RNA from 8 cell lines profiled on 5-bp resolution tiling arrays	Human	61% of novel transcribed fragments represent transcripts from both strands of the genome; 50% of intronic sequences are antisense	NA	NA	(Cheng et al. 2005; Kapranov et al. 2005)
1,947 S-AS pairs profiled on oligoarrays using poly(A)+/- RNA	Mouse	NA	c/c*: 48% c/nc: 43% nc/nc: 8%	NA	(Kiyosawa et al. 2005)

C

Dataset	Organism	Antisense transcription	Coding status	Topology			Reference
Short SAGE library of U937 cell line	Human	6.3% of expressed mRNAs had antisense tags	NA	NA			(Quere et al. 2004)
8 LongSAGE libraries, one tissue	Mouse	3.5% of all tags were antisense to genes	NA	NA			(Wahl et al. 2005)
72 LongSAGE libraries, multiple tissues (Mouse Atlas of Gene Expression)	Mouse	Approximately one third of mapped LongSAGE tags are antisense	NA	NA			(Siddiqui et al. 2005)
43,553 TUs from Fantom3; GenBank cDNAs, ESTs, CAGE, GIS	Mouse	Strong evidence: 29% Good evidence: 51% Weak evidence: 72%	c/c: 37% SA, 26% NOB c/nc: 55% SA, 59% NOB nc/nc: 8% SA, 14% NOB	D 36%	C 34%	F 30%	(Riken Genome Exploration Research et al. 2005)

2. Next generation tag sequencing for cancer gene expression profiling[2]

Author contributions

T.Z, H.M., Y.Z., and M.H. were involved in the design of the Tag-seq method, and in creation of both Tag-seq and LongSAGE libraries (sections 2.2.1, 2.4.1, 2.4.2). S.J. and M.A.M. supervised the project and contributed design concepts and comments throughout. The data filtering algorithm was designed and implemented by A.D. (sections 2.2.2, 2.4.2, 2.4.3), who also conducted inter-plaform correlations measures (sections 2.2.4.1, 2.2.4.2, 2.4.6). S.D. processed RNA-seq data. R.M. wrote scripts used in section 2.4.7, was involved in original manuscript planning, and contributed text to the manuscript. I (A.S.M.) performed the majority of the computational experiments in the text, designed and performed the cancer-related analyses (sections 2.2.2-2.2.8 and 2.4.4, 2.4.5, 2.4.7-2.4.9), created figures and tables describing the results, and wrote the manuscript.

2.1 Introduction

A key first step in understanding cellular processes is a quantitative representation of gene expression profiles, including those relevant to cancer. As part of the Cancer Genome Anatomy Project (CGAP), the gene expression profiles of a wide variety of cancer tissues and cells were measured using LongSAGE libraries, created and sequenced using conventional Sanger sequencing methods (Lal et al. 1999). Prior to completion of the CGAP project, the advent of new massively parallel sequencing technologies made feasible an improvement in the efficiency and sensitivity with which tag-based gene expression could be measured. We thus sought to develop and apply a next-generation sequencing approach for tag-based gene expression profiling to complete the CGAP database.

---

[2] A version of this chapter has been published. Morrissy AS, Morin RD, Delaney A, Zeng T, McDonald H, Jones S, Zhao Y, Hirst M, and Marra MA. 2009. Next generation tag sequencing for cancer gene expression profiling. *Genome Res*, **19**:1825-35

A portion of this chapter has been published. Morrissy AS, Zhang Y, Delaney A, Asano J, Dhalla N, Li I, McDonald H, Pandoh P, Prabhu A-L, Tam A, Hirst M, and Marra MA. 2010. Digital Gene Expression by Tag Sequencing on the Illumina Genome Analyzer. *Current Protocols in Human Genetics*. **Supp 65. Unit 11.11.** John Wiley and Sons, Hoboken, NJ.

Several recently developed sequencing technologies, such as Roche Applied Sciences' 454 pyrosequencing platform (Margulies et al. 2005), the Illumina Genome Analyzer (Bentley 2006), and Applied Biosystems SOLiD platform (Valouev et al. 2008), offer massively parallel production of short reads. Using these technologies, thousands to millions of isolated and amplified DNA molecules can be attached to a solid surface (such as a flowcell or microbeads), and sequenced in parallel. Such technologies offer up to two orders of magnitude increase in per base cost-efficiency compared to capillary sequencing (von Bubnoff 2008). These platforms have made feasible previously cost-prohibitive projects such as genome re-sequencing (Green et al. 2006; Bentley et al. 2008; Ley et al. 2008; Wang et al. 2008b), and deep transcriptome and non-coding RNA sequencing (Nielsen et al. 2006; Weber et al. 2007; Marioni et al. 2008; Morin et al. 2008; Rosenkranz et al. 2008), as well as genome-wide protein binding-site surveys (Chip-seq) (Jothi et al. 2008; Wederell et al. 2008).

The high-throughput methods preceding the massively parallel sequencing approaches mentioned above are diverse, but can generally be classified either as sequence-based or hybridization-based. The former are often termed 'digital' because they reflect the number of individual observations of a transcript, while the latter, typically in the form of microarrays, are termed 'analog' as they provide a surrogate hybridization-based measure of individual transcript abundance. Digital gene expression profiling using ESTs (Adams et al. 1991; Hillier et al. 1996) was cost-restrictive and more cost efficient tag-based techniques such as serial analysis of gene expression (SAGE) were developed (Velculescu et al. 1995). Despite increases in cost-efficiency compared to EST profiling, the expense and specialized facilities required for high-throughput capillary sequencing prevented SAGE from becoming as widespread as its microarray counterparts.

Our goal was to implement a tag sequencing protocol on the Illumina platform, analogous to LongSAGE (Saha et al. 2002), and to use this protocol to measure transcript abundance in human cancers. The Illumina (Bentley et al. 2008) sequence-by-synthesis technology currently offers approximately 80 million reads (10 million reads per lane; 8-lane flow cell)[3] from a single run of the instrument. This makes possible gene expression profiling experiments with much improved dynamic range and considerable cost savings

[3] As of 2010, the Illumina instrument generates approximately 62 million reads per lane.

compared to capillary sequencing of LongSAGE libraries. Our approach, called Tag-seq, generates 21 base-pair tags, generally from the 3' ends of transcripts. The method is similar to the LongSAGE approach, but forgoes the need for ditag production, concatenation, and cloning. Deep sequencing of tags is achieved using only a single lane of a flow cell, and typical yields are in the range of 5 – 10 million sequences.

Compared to conventional microarrays, Tag-seq should not suffer from cross-hybridization of related sequences, and in principle offers dynamic range limited only by sequencing depth. Compared to RNA-seq, Tag-seq performs comparably in terms of gene discovery and dynamic range. While Tag-seq does not provide information regarding the internal structure of transcripts, it can distinguish between transcripts originating from either of the DNA strands. There are advantages in using a strand-specific gene expression platform, for example to measure the prevalent antisense transcription in the human genomes (Riken Genome Exploration Research et al. 2005). Here, we conduct an analysis of Tag-seq data from the CGAP collection to illustrate the utility of the method in addressing questions of relevance to cancer biology.

2.2 Results

2.2.1 Data generation

The Tag-seq protocol (Morrissy et al. 2010b) is similar to the LongSAGE approach (Saha et al. 2002), in which a restriction endonuclease (NlaIII) cleaves each individual transcript in a sample, and a type II restriction endonuclease (MmeI) is used to generate a 21bp tag from the 3'-most NlaIII site. In LongSAGE, tags from individual transcripts are ligated together to form ditags that are concatenated, cloned, and sequenced using capillary sequencing. The Tag-seq method, in contrast, forgoes ditag production and concatenation, and allows the direct sequencing of tags using massively parallel sequencing on the Illumina Genome Analyzer ((Morrissy et al. 2010b), **Fig. 2.1**). Typically, a Tag-seq library is sequenced to a depth of 10 million tags, which represents an increase of 2 orders of magnitude over the sequencing depth of a typical LongSAGE library. Our expectation was that the added depth of the Tag-seq method would improve representation of important low-abundance transcripts at the limits of or beyond LongSAGE sensitivity.

We used the Tag-seq platform to complete the CGAP digital gene expression profiling project, by generating 35 libraries from cancer and normal tissue samples. To assess the similarities between the new Tag-seq data and the existing LongSAGE data, we compared the data from these 35 libraries to those from 77 LongSAGE libraries. In total, we produced two meta-libraries, one containing 6.9 million LongSAGE tags from the 77 libraries (1.1 million distinct tag sequences), and one containing 170 million Tag-seq tags from the 35 quality filtered libraries (4.0 million distinct tag sequences). These libraries are publicly available as part of the Cancer Genome Anatomy Project (CGAP) collection ((Lal et al. 1999); **Appendix A**). The CGAP libraries also included two replicate libraries, one Tag-seq library and one LongSAGE library, which were created from the same human embryonic stem cell (hESC) RNA source.

2.2.2 Data filtering

To ensure that we analyzed high quality data in the Tag-seq libraries, we removed potentially erroneous tags using a novel filtering algorithm (Methods). Briefly, tags were removed if they occurred once (singletons), or if they differed by one basepair from more highly expressed tags (one-offs) unless they mapped to the genome or transcriptome. On average, 22.1% of filtered tags could be mapped to Ensembl transcripts, while only 1.2% of tags removed by the filter could be mapped to transcripts. While filtered tag sequences comprised an average of 7.5% of all tag sequences, their abundance corresponded to an average of 56.0% of the total library size, and they identified over 97.5% of the total number of genes detected by all tags.

2.2.3 Effect of library depth on tag sequence diversity and abundance

By comparing the Tag-seq and LongSAGE meta-libraries, we sought to first determine whether differences in Tag-seq and LongSAGE protocols resulted in any significant bias in tag or gene representation. As expected, we found a significant overlap between these meta-libraries**,** with over 300,000 unique tag sequences detected using both methods. On average, these commonly detected tag sequences were expressed in a larger proportion of Tag-seq libraries than LongSAGE libraries, and had 17-fold higher expression in Tag-seq libraries (**Table 2.1**). A large number of tag sequences were detected by only one method; in general, these were expressed at lower levels than those tag sequences found by both methods, and in fewer libraries. The 3 million tag sequences detected only by

Tag-seq were on average 1/16 the abundance of the tags detected in common by both methods (absolute counts, **Table 2.1**), and therefore were likely undetectable in the LongSAGE libraries due to their comparatively shallow sequencing depth. Thousands of Tag-seq tag sequences did not map to any unique or repetitive sites in the genome or the transcriptome. These may indicate the presence of either novel transcripts, or novel isoforms of annotated genes that lead to the creation of novel tag sequences spanning splice sites (80,875 Tag-seq tag-sequences and 63,166 LongSAGE tag sequences expressed over counts of 10; **Fig. 2.2C**).

Nearly a third of the tags detected in both meta-libraries mapped to 21,638 genes. A small proportion of tag sequences found solely in LongSAGE (8.1%) or Tag-seq (3.5%) mapped to Ensembl genes (**Table 2.1**). Although in general the tag sequences found only by Tag-seq had expression levels below those detectable by LongSAGE, the 741 genes found only in Tag-seq had an average expression level higher than that for the genes found in common. They are therefore likely to be genes specific to tissues not profiled by LongSAGE. With the exception of the hESC replicate libraries (section **2.2.1**), all LongSAGE and Tag-seq libraries represented diverse tissues, although the greater number of LongSAGE libraries doubled the diversity of tissues profiled by LongSAGE. The 430 genes found only by LongSAGE were on average less frequently expressed than genes detected by both methods, and may represent genes specific to tissues profiled using LongSAGE.

We next investigated the effect of depth on gene representation by comparing the Tag-seq and LongSAGE replicate libraries created from the same hESC RNA sample. The Tag-seq replicate (library id 'hs0238') had a total of 293,179 tag sequences (filtered tags only), of which 40,149 (13.7%) mapped to Ensembl genes, either in introns, exons, or on the opposite strand. The LongSAGE replicate (library id '1313' in Table S1) had a total of 19,998 tag sequences, of which 13,983 (69.9%) mapped to Ensembl genes. The LongSAGE tag sequences mapped to 7,055 genes and the Tag-seq tag sequences mapped to 11,165 genes, which included 93.5% of the genes found by LongSAGE. Thus, added depth improved gene detection in this tissue 1.6-fold. Since each tag sequence mapping to a gene can represent an individual transcript isoform (Siddiqui et al. 2005), we analyzed the average expression of all transcript isoforms. The transcripts of the 6.5% of

genes only found by LongSAGE were expressed at low levels (average of 4.0 counts), and may be under-represented in the Tag-seq library due to variability in the replicate library creation. The detection of transcription factors was 1.8-fold greater, with 429 TFs detected by LongSAGE, and 799 TFs detected by Tag-seq. The average expression of the 393 TFs detected in common was higher (69.8 in the Tag-seq replicate, 6.8 in the LongSAGE replicate), than that of the 36 TFs detected only in LongSAGE (5.9), and the 406 TFs detected only by Tag-seq (26.7).

To determine whether these additional genes found by Tag-seq were functionally different than those found by both methods, we conducted an assessment of GO categories overrepresented in the Tag-seq versus the LongSAGE replicate (Ashburner et al. 2000). The most significantly overrepresented terms in this tissue were found by both methods. Thus, increased sequencing depth resulted in identification of additional genes that belonged to the same functional categories (data not shown).

We next asked whether a Tag-seq library unambiguously identified a larger number of genes on average than a standard LongSAGE library. We performed a sampling simulation to estimate the number of genes represented by different 'depths' of sequencing in each Tag-seq and LongSAGE library. Sampling up to 300,000 tags from individual LongSAGE libraries resulted in detection of up to 10,000 genes (**Fig. 2.2A**). Quality filtered Tag-seq libraries sampled at depths of up to 10 million tags detected up to 13,000 genes. This suggested that the added depth provided by the Tag-seq approach results in a more comprehensive interrogation of gene expression profiles, with 48.3% and 36.3% of expressed genes detectable at depths greater than those of a typical (100,000 tags) or large (200,000 tags) LongSAGE library, respectively. At every sampling depth level greater than 1 million tags in Tag-seq, the rate of gene detection was reduced (**Fig. 2.2B**).

2.2.4 Differences in gene abundance between Tag-seq and other gene expression platforms

2.2.4.1 Tag-seq vs LongSAGE and SAGELite

Having established that the measured sampling depth of Tag-seq improved gene discovery, we evaluated the concordance of tag abundance between the two methods, by re-analyzing the Tag-seq and LongSAGE replicate hESC libraries. The LongSAGE

replicate had a total of 272,465 tags, while the Tag-seq replicate had a total of 3,636,083 quality filtered tags. Tags expressed in common between these libraries had a Pearson coefficient of 0.60 (**Fig. 2.3**). We analyzed another set of replicate Tag-seq and LongSAGE libraries created from the same mouse RNA (Methods), and found they had a Pearson correlation of 0.64. This was comparable to the correlation between the LongSAGE library and a technical replicate generated with the SAGELite protocol (0.64). SAGELite is a variant of LongSAGE used to create libraries from samples that are too small to yield sufficient amounts of mRNA for standard LongSAGE library construction (Peters et al. 1999). We observed a lower Pearson coefficient between the Tag-seq technical replicate and the SAGELite replicate (0.43), indicating these methods have different biases relative to LongSAGE.

2.2.4.2 Tag-seq vs Affymetrix

We generated Pearson correlations between three non-CGAP Tag-seq libraries and their respective technical replicates analyzed on Affymetrix exon arrays. Correlations were calculated for expressed tags that represented known transcripts and mapped uniquely or not at all to the genome, and their corresponding Affymetrix probes. Pearson coefficients for the three technical replicates were very similar to each other (0.59, 0.60, and 0.61), and to that of Tag-seq and LongSAGE replicates.

To determine whether the Tag-seq platform performed better in quantifying the dynamic range of expressed genes, we analyzed one of the three Affymetrix:Tag-seq replicates. We binned the 10,152 genes detected in common between the two platforms by expression level into 10 bins, from least to most highly expressed (**Fig. 2.3E**). In the Affymetrix replicate, the genes in bins 1-4 were indistinguishable from background noise. Thus, genes with measurable expression above background in Affymetrix were contained in bins 5-10 only. In contrast, genes with measurable expression above background in Tag-seq were well separated among the 10 bins. Between Affymetrix bin5 and bin10, genes had a fold-change of 78.7, while between Tag-seq bin1 and bin10, genes had a fold change of 1,018.5, nearly 13 times higher. Between Tag-seq bins 2 and 10, the fold change was 407.6, over 5 times higher.

The log transformed minimum, mean, and maximum expression was next calculated for the tag sequences in each library. The range between the minimum and the maximum,

and between the mean and maximum values was computed (**Table 2.4A**). Tag-seq counts and Affy intensities spanned the same range (2.2-11.0 and 0.0-10.9, respectively, averaged over three replicate libraries), however the range between the average gene expression and the max gene expression was double in Tag-seq vs RNA-seq (8.2 vs 4.9), showing that genes expressed above average have twice the log-transformed dynamic range in Tag-seq.

2.2.4.3 Tag-seq vs RNA-seq

We also analyzed a pair of replicate RNA-seq / Tag-seq libraries created from the same RNA source (Methods), and found a high concordance in transcript identification between the methods. A total of 8,050 transcripts were found by both methods (**Table 2.4B**), representing 94.4% of all 8,528 transcripts found by RNA-seq, and 96.2% of all 8,366 transcripts found by Tag-seq (Pearson correlation of gene abundance: 0.54). Commonly expressed genes were also highly concordant in terms of dynamic range (Pearson correlation of 0.97; see Methods).

The library construction approach used to make Illumina libraries does not currently distinguish between reads derived from opposing DNA strands, and RNA-seq reads were therefore not able to discriminate between sense and antisense transcription. For nearly a third (29.5%) of the genes detected by both methods in this replicate library set the Tag-seq replicate detected expression on the antisense strand (**Table 2.4B**). In the case of 613 loci detected by both methods, the Tag-seq reads clearly show that expression arises solely from the antisense strand. At these loci, correlations between gene expression levels measured by Tag-seq vs RNA-seq (0.50) were the same as those at loci with sense expression in both technologies (0.54).

2.2.5 GC-content bias

We next investigated whether there was any detectable bias in the sequence composition of tags profiled by the Tag-seq and LongSAGE platforms. The GC-bias of a platform can be calculated by comparing the number of standard deviations by which the observed bias in an individual library deviates from that of the expected bias ((Siddiqui et al. 2006); Methods). We found that Tag-seq libraries were significantly more AT-rich than LongSAGE libraries (**Fig. 2.4A**). As previously observed, LongSAGE libraries had a

weak GC-bias (-3.51 +/- 8.08), while Tag-seq libraries had a stronger AT-bias (12.99 +/- 5.39), comparable to that of the Affymetrix platform (HGU 133 GeneChip; ((Siddiqui et al. 2006)). As observed for Affymetrix, this bias decreased in parallel with increasing expression level, such that highly expressed Tag-seq sequences were significantly less biased (all filtered tag sequences vs those expressed over counts of 500, $P = 2.1 \times 10^{-10}$, T-test). This suggests that as sequencing depth increases in sequencing-based technologies, a distinct class of genes with increasing AT content is detected. We tested whether this was the case in Tag-seq by comparing the GC-content of the genes with high vs low frequency tags, and found that genes expressed at or below 100 tag counts were significantly more AT-rich than genes expressed at or above 1,500 tag counts ($P = 2.8 \times 10^{-4}$, T-test , **Fig. 2.4B**). This was true of gene sequences that included introns, but not of cDNA sequences (data not shown), indicating that the AT-content of the genomic regions in which these genes were encoded was correlated to their expression level. In LongSAGE, bias also decreased with increasing expression level, such that tag sequences expressed over 20 and over 100 counts become significantly less biased (all tag sequences vs those expressed over counts of 100, $P = 1.9 \times 10^{-3}$, T-test). This trend was also correlated to the GC-content level of the genes to which LongSAGE tags mapped to, indicating that the source for these observations was also biological in nature rather than a technical artifact (**Fig. 2.4B**).

Next we determined the extent to which tag sequence representation was biased in Tag-seq versus LongSAGE, by re-analyzing the hESC replicate libraries made from the same RNA source. Tag sequences detected solely by LongSAGE had a greater GC-content than those detected solely by Tag-seq (0.50 vs 0.39), however, both sets of tag sequences were on average very infrequently expressed (**Fig. 2.5A**). In contrast, the 13,161 tag sequences detected by both methods were highly expressed and had an intermediate GC-content (0.43) that was nearly identical to the average GC-content of all Ensembl transcript tag sequences (0.42). We looked at whether the correlation of expression of these common tag sequences differed as a function of tag GC-content. We divided the tags into four bins representing increasing proportions of tag GC-content (bin1: 0%-25%, bin2: 25%-45%, bin3: 45%-65%, bin4: 65%-100%), and found that the Pearson correlation changed as a function of GC-content, with AT-rich tags having the lowest correlation between methods (**Fig. 2.5B**).

We investigated the cause of the decreased correlation between AT-rich tag sequences in the two methods, and found a relationship between tag abundance and tag GC-content. In LongSAGE we observed a positive correlation between tag abundance and GC-content for the first 3 bins (bin1 vs bin2 $P = 1.6 \times 10^{-3}$, bin2 vs bin3 $P = 1.4 \times 10^{-3}$, T-test). In contrast, the abundance of the same tag sequences in the Tag-seq replicate did not correlate with GC-content, with the exception of the most GC-rich bin (bin3 vs bin4 $P = 9.4 \times 10^{-8}$; **Fig. 2.5C**). This relationship between GC-content and tag abundance held for all Tag-seq and all LongSAGE libraries (data not shown).

2.2.6 Improved representation of low abundance LongSAGE transcripts in Tag-seq libraries

Given the increased depth of Tag-seq libraries, we expected to observe increased numbers of tags for transcripts at the limit of detection in LongSAGE (Siddiqui et al. 2005). Two such tag categories include antisense and intronic tags. Antisense tags originate from transcripts that are transcribed from the opposite strand (**Fig. 2.6**), while intronic tags may represent unannotated exons and UTRs within known genes (Saha et al. 2002), previously unannotated sequences transcribed from introns, such as embedded genes (eg. HA_003240, (Hirst et al. 2007)) or miRNA genes (Kim 2005). Another class of generally low abundance transcripts of biological interest consists of transcription factors (TFs). To investigate the expression levels of TFs in Tag-seq and LongSAGE libraries, we downloaded the set of 2,890 human genes that encoded DNA-binding domains (DBD, http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?About), which should include all TFs, and searched for their presence in the CGAP libraries.

We enumerated tag sequences that mapped in the sense orientation to TF exons, antisense to known genes, and sense to gene introns, in each library, at increasing thresholds of expression. Overall, an average Tag-seq library detected 1.7 times as many TF genes as a LongSAGE library (849 vs 504), 6.3 times as many genes with AS tags (4,999 vs 795), and 2.8 times more genes with intronic tags (7,651 vs 2,752). The majority of genes found by Tag-seq were at expression levels below those detectable in existing LongSAGE libraries (**Fig. 2.7A-C**).

We confirmed the relationship between sequencing depth and the diversity and abundance of intronic and antisense tags by analyzing the Tag-seq and LongSAGE hESC

replicate libraries. To ensure that the relationship between tag sequence diversity and tag abundance was due to no other factors except depth, we generated an *in silico* library of 272,465 randomly sub-sampled tags from the Tag-seq replicate. The in silico library, hereafter referred to as sub_Tag-seq, theoretically represents a random sample of the most highly expressed tags in the Tag-seq replicate, and should therefore be very similar to the LongSAGE replicate. We found that sub_Tag-seq was moderately correlated with the LongSAGE replicate (Pearson correlation of 0.6), with most of the variation coming from low frequency tags (data not shown). Any differences in the abundance of intronic and antisense tags in sub_Tag-seq library relative to the Tag-seq library would most likely be due to decreased depth.

A comparison of the Tag-seq replicate, sub_Tag-seq, and the LongSAGE replicate supports the described increase in the diversity of intronic and antisense tags in deeper libraries. We compared the proportion of tag sequences in each library that mapped either to exons, introns, or to the antisense strand of Ensembl genes (**Fig. 2.8A**). In the Tag-seq replicate, the most abundant categories of mapped tag sequences were exonic tags (47.8%), followed by antisense tags (32.1%), and intronic tags (20.6%). In contrast, the LongSAGE replicate was far more likely to detect tags mapping to exons (73.0%) than antisense (23.4%) or intronic tags (6.2%). Thus, the Tag-seq replicate is enriched in antisense and intronic tag sequences; this enrichment is not observable at sampling depths less than 300,000 tags, since the tags in sub_Tag-seq library mapped in proportions similar to those of the LongSAGE replicate (differences were not significant). These observations held when comparing all Tag-seq to all LongSAGE libraries (data not shown), indicating that low-frequency antisense and intronic tags were present in all the profiled human tissues, and were not specific to hESCs. The altered proportion of antisense, intronic, and exonic tag sequences was highly significant (T-test between Tag-seq and LongSAGE tag sequence proportions: antisense $P = 6.2 \times 10^{-5}$, intronic $P = 1.0 \times 10^{-10}$, exonic $P = 1.6 \times 10^{-24}$).

Interestingly, the abundance of exonic, intronic, and antisense tag sequences was almost identical between methods (**Fig. 2.8B**). Thus, in both methods exonic tags were the most abundantly expressed (~80%), followed by antisense tags (~20%), and intronic tags (0.1%).

The additional depth in Tag-seq had a dramatic effect on the dynamic range of expression of moderate to abundantly expressed tags, which could be detected by both methods. On average, exonic tag sequences were detected at frequencies 12.7-fold higher in the Tag-seq versus the LongSAGE replicate, and antisense and intronic tag sequences were detected at levels 13.4 and 14.4-fold higher (**Fig. 2.8C**). The range of expression was an order of magnitude higher in Tag-seq versus LongSAGE, indicating a significantly greater dynamic range of expression.

2.2.7 Sense-antisense transcripts in cancer libraries

Having assessed the technical differences between the LongSAGE and Tag-seq protocols, we undertook a biological analysis of the CGAP library collection. We first analyzed the antisense tags with a focus on their differential expression in libraries representing cancerous and normal tissue samples. Previous studies have shown that the ratio of sense to antisense transcripts changes between normal and malignant tissue samples (Chen et al. 2005), and that antisense transcripts can be implicated in disease processes (Tufarelli et al. 2003; Reis et al. 2004). Our goal was to highlight the potential of the Tag-seq approach to identify known and novel antisense transcripts whose expression ratios changed significantly with respect to the sense gene, between normal and diseased states, between different stages of disease progression, or between cancer subtypes.

To achieve this, libraries were first grouped by tissue into 15 groups (**Appendix A**; Methods). Libraries belonging to each tissue were segregated into groups representing normal and cancerous samples, and when possible, were further segregated into cancer stages (pre-cancerous samples versus malignant for instance; **Appendix B).** The ratio of sense to antisense transcription between each of the tissue groups was assessed at every relevant locus; either using pairs of sense tags mapping to known SAS gene pairs, or using sense tags mapping to single genes with novel antisense transcription (novel SAS) (**Fig. 2.6**).

Altered expression ratios between 389 known SAS gene pairs and between 2,195 novel SAS pairs were found in the 15 tissue groups. Random assignment of tags to genes showed that known SAS genes were, on average, 55 times more likely to have ratio changes than would be expected by chance, while novel SAS genes were 17.5 times

more likely than expected by chance, suggesting a higher rate of false positives in the ratio changes of these pairs. We developed a normalization protocol to identify pairs with large expression ratio changes (Methods), and to ensure higher ranking of highly expressed gene pairs and of those pairs with lower variance in their ratios. Overall, tissues comprised solely of Tag-seq or LongSAGE libraries had equivalent numbers of gene pairs with ratio changes. Since the tissues profiled by the different methods were distinct, we could make no a priori predictions regarding the number of gene pairs with different ratios found by Tag-seq or LongSAGE. By definition, the genes targeted by this analysis are moderately to highly expressed, and could be found by both methods. Thus, in the absence of Tag-seq and LongSAGE replicates for a whole tissue, we conclude that both methods are capable of finding gene pairs whose abundance ratios change between cancerous and normal samples, and which therefore may be differentially regulated in cancer versus normal tissues.

To determine whether there was an enrichment of biological categories in these genes, we conducted a functional annotation clustering analysis (Dennis et al. 2003; Huang et al. 2009). In this analysis, annotations (such as gene ontology (GO) terms; (Ashburner et al. 2000)) that share common genes are more likely to be grouped together. We found that genes with extreme ratio changes (in the top 10%) were highly enriched in GO terms relating to the regulation of developmental processes, to the regulation of cell death, and to cell proliferation, terms which are relevant to cancer biology (data not shown).

To further evaluate the biological relevance of these pairs, we enumerated the number of Cancer Gene Census genes in the dataset (Futreal et al. 2004). This is a catalog of genes with mutations that have been causally implicated in multiple cancers. Of the total 312 cancer census genes, expression was detected in the CGAP dataset for 300. Interestingly, over one quarter of these genes (72 novel and 6 known SAS) were also found to have significant ratio changes between normal and cancerous libraries in the studied tissues (**Table 2.2; Appendix C**). The pairs with ratio differences in the top 10% of the range of differences were identified, revealing a total of 30 of the cancer census genes remaining in this shortlist (27 novel and 3 known SAS). Thus, 38% of the cancer genes were in the top 10% of differentially expressed genes with extreme ratio changes between cancer and normal tissues, which is a significant enrichment ($P < 7.0 \times 10^{-4}$, Chi-square test).

2.2.8 Transcript isoforms in cancer libraries

Differential expression of transcript isoforms was analyzed in 4,237 genes with multiple expressed tags, since these tags potentially represent alternative 3' polyadenylation sites (Siddiqui et al. 2005). A total of 1,957 of these genes had tag pairs whose ratio of expression changed between libraries grouped by disease state (eg. cancerous vs normal). For 1,304 (66.6%) of these genes, the sequence bounded by the two tags harboured predicted miRNA targeting sites (Grimson et al. 2007), suggesting that miRNAs may regulate isoform expression in one of the two states (Hirst et al. 2007). The proportion of miRNA-targeted genes in this list was nearly three times greater than the proportion of miRNA-targeted genes in the human genome (22.0%, $P < 2.2 \times 10^{-16}$, Chi-square test; **Table 2.3**). Of the 772 genes with transcript pairs that had the 10% most extreme expression ratio changes, we found an additional enrichment of transcripts harbouring miRNA targeting sites (72.5%; **Table 3**). For 33.1% of these genes, the longer isoform was consistently more abundant in cancers; for 41.0% of these genes, the shorter isoform was consistently more abundant in cancer; for the remaining 26.9% of genes, either isoform was more abundant in cancer in some sample.

We found 93 miRNA targeting sites with enriched frequencies in the set of genes with the top 10% most extreme expression isoform ratio changes (versus the frequencies in the set of all genes with isoform ratio changes, $P < 0.05$, hypergeometric distribution test; **Appendix D**). A closer look at the most enriched sites showed that these miRNAs have been previously observed to have altered expression in cancers (eg. miR-124 in glioblastoma multiforme, (Silber et al. 2008); miR-181 and miR-15/16 in B-cell chronic lymphocytic leukemia, (Calin et al. 2002; Pekarsky et al. 2006); miR-224 in thyroid tumors and in hepatocellular carcinoma, (Nikiforova et al. 2008; Wang et al. 2008c)).

2.3 Discussion

To complete the CGAP digital gene expression profiling project, we developed Tag-seq as an efficient and cost effective alternative to LongSAGE. Tag-seq library construction is similar to the LongSAGE protocol, but sequencing employs Illumina's massively parallel sequencing by synthesis protocol in place of conventional Sanger sequencing. Every read in a sequenced Tag-seq library represents a 17-bp sequence tag adjacent to

the 3'most NlaIII site of an individual transcript, and therefore represents a digital count of that transcript.

Relative to another Illumina-based transcript profiling technology, RNA-seq (Marioni et al. 2008; Rosenkranz et al. 2008), Tag-seq performs comparably in terms of gene discovery and measured dynamic range. For gene expression profiling experiments where accurate profiling of transcripts from both strands of the genome is required, Tag-seq data are superior, since unlike current applications of RNA-seq (see section 5.1), it allows discrimination of sense and antisense transcripts. Sense and antisense genes are encoded on the opposite strands of the same genomic locus, and yield transcripts that have sequence complementarity. Their genomic arrangement and sequence complementarity increases the likelihood that their regulation is affected by common factors (such as chromatin state) and their relative expression (such as transcriptional interference), at both the transcriptional and post-transcriptional level (Lavorgna et al. 2004; Dahary et al. 2005). To date, antisense transcripts have been observed for up to 72% of the mammalian transcriptome in datasets generated by both sequence-based and hybridization-based methods (Riken Genome Exploration Research et al. 2005). Given the high prevalence of antisense transcription in the mammalian genome, and the link between antisense transcripts and disease (Tufarelli et al. 2003; Reis et al. 2004), Tag-seq was well suited to the study of cancer-relevant gene expression in the context of the CGAP project. We found known and novel SAS gene pairs for which the ratio of expression changed significantly between cancer subtypes or between cancer and normal states. These were enriched in known cancer-related genes, supporting a role for antisense transcription in cancer biology. For instance, we found evidence for antisense transcription at the BCL6 locus, which encodes an oncogene that is known to be involved in lymphomas (Ye et al. 1997). Antisense ESTs have previously been observed at this locus, lending support to our observations of antisense transcription (**Fig. 2.9**). The number of antisense tags at this locus was significantly increased in the subset of libraries from grade II carcinoma epithelium and associated myofibroblast samples, leading to a reduced sense-to-antisense ratio in those samples. These libraries represented cell types sampled from one breast cancer patient, implicating the relationship between BCL6 and its antisense transcript in the biology of this individual breast cancer. While carcinoma associated myofibroblasts are not necessarily cancer cells *per se*, they have epigenetic

alterations similar to those seen in malignant carcinoma epithelium, and are globally hypomethylated (Jiang et al. 2008). One plausible explanation for the increase in antisense expression at this locus is increased hypomethylation at CpG islands downstream of the BCL6 gene (**Fig. 2.9**).

While Tag-seq is able to distinguish transcript strand of origin, it only provides limited information regarding transcript structure. Thus, to gather data on expressed transcript isoforms, exon arrays or RNA-seq would be the more suitable technologies. However, Tag-seq is still informative on the expression of the subset of gene isoforms that lead to a different 3' NlaIII tag sequence as a consequence of alternative 3'-end formation. We were able to analyze over 4,200 genes with such transcript isoforms and expression in CGAP, and to find differential expression of isoforms between cancer and normal states. Intriguingly, we found an enrichment of transcripts harbouring miRNA targeting sites in the sequence unique to one of two differentially expressed isoforms (Hirst et al. 2007; Ghosh et al. 2008), implicating their regulation in cancer biology.

Compared to Affymetrix microarrays, Tag-seq is capable of de novo gene discovery without the requirement of genome-wide probe design, does not suffer from cross-hybridization of related sequences, and achieves essentially unlimited dynamic range simply by increasing sequencing depth. At the current level of sampling (~10 million tags), genes detected by Tag-seq had a 13-fold greater measurable fold change than the same genes detected by Affymetrix.

Relative to LongSAGE, the additional depth of sampling provided by Tag-seq led to a greater number of genes identified in a given tissue, and improved the measurable dynamic range of those genes. One other report has thus far shown that Tag-seq surpasses LongSAGE in sequencing depth (Hanriot et al. 2008). We extend these findings by reporting for the first time that with increasing depth, Tag-seq also allowed detection of a distinct subset of transcriptome space, enriched in AT-rich genes, intronic tags, antisense tags, and novel intergenic tags. The enhanced detection of low-frequency AT-rich tag sequences in Tag-seq was similar to previous observations made in Affymetrix arrays (Siddiqui et al. 2006), although the detection of AT-rich sequences was in that case interpreted as a technological bias. These new results suggest that this AT-rich class of tag sequences do not represent technical bias in either method, but rather

a biological difference in the types of transcripts present at lower frequencies, which is detectable using both sequencing-based and hybridization-based technologies. The depth of sampling achieved by LongSAGE is not large enough to detect this subset of the transcriptome. Furthermore, we found that Tag-seq has less GC-bias, leading to a more accurate interpretation of the abundance of tags spanning the range of GC-content.

Overall, Tag-seq identifies more genes than LongSAGE, detects a greater dynamic range of expression, and thus allows differential expression analysis for a greater range of transcripts. Tag-seq libraries provide an excellent resource for the discovery of known and novel transcripts with expression changes relevant to disease processes, and highlight the applicability of next generation tag sequencing to gene expression profiling.

2.4 Methods

2.4.1 Tag-seq library construction

All libraries were constructed using one of two protocols: Tag-seq or Tag-seqLite. Tag-seq is a variant of LongSAGE as described (Siddiqui et al. 2005; Khattra et al. 2007), with modifications forgoing the requisite production of ditags and concatemers and allowing direct sequencing on the Illumina Genome Analyzer (**Fig. 2.1**). Typically 500-2000ng of DNase I treated total RNA was used in Tag-seq library construction, and 4-50ng in Tag-seqLite library construction (Morrissy et al. 2010b).

Tag sequences as long as 26bp are possible (SuperSAGE; (Matsumura et al. 2005)) by using EcoP15I as a tagging enzyme, but there are drawbacks to using this enzyme, in return for marginal gains in tag-to-gene mapping.

EcoP15I is a Type III enzyme, requiring two oppositely oriented binding sites for cleavage. To use this enzyme with the current Tag-seq protocol, EcoP15I binding sites would need to be added in the 5' adapter (at the NlaIII site) and the 3' adapter (part of the bead-bound Oligo-dT primer). One issue is that cleavage can occur at either one of the binding sites, leading to a different tag sequence for the same transcript in different libraries, which would confound differential gene expression analysis and make comparison to LongSAGE impractical. Another issue is that the efficiency of EcoP15I cleavage changes with increasing sequence length between the two binding sites (the 3'most NlaIII site and the end of each mRNA), thus further affecting the ability to collect

quantitative gene expression measurements (http://www.neb.com/nebecomm/products/productR0646.asp). Additional aspects of the cleavage reaction that affect cutting efficiency and specificity (ie. which of the two recognition sites is used for cleavage, whether base composition biases this choice, which cofactors affect cutting efficiency, etc) are still under study (Moncke-Buchner et al. 2009).

Even if the specific conditions required for efficient and specific EcoP15I cleavage steps were known, the cost of changing the reagents used in the Tag-seq protocol would become an issue, since the Illumina Oligo-dT beads can not be used with EcoP15I without modification. The modification required would be the addition of linker sequence (including the EcoP15I cutting site) on the bead-bound polyT primer. Ordering custom-made polyT-beads would increase the time and cost required for library construction, and decrease the wide accessibility of the method to other investigators.

MmeI is therefore a better tagging enzyme choice than EcoP15I for tag-based sequencing (Saha et al. 2002), due to its simpler cleavage parameters and adequate tag-length.

2.4.2 Tag extraction

Sequencing of a Tag-seq amplicon starts at the first base following the Adapter A sequence. Thus, the first 17 to 18 bases of a read are the transcript-derived tag sequence, and the remaining bases are the Adapter B sequence. As expected, 99% of adapters found in a Tag-seq library occur in positions 18 and 19 of the read. The "Raw" Tag-seq library is then constructed by truncating all reads at length 17.

2.4.3 Tag-seq filtering

In the SSOOHE (Sans Singletons and One-Offs of Highly Expressed tags) filtering algorithm, tag sequences are removed if they are only observed once (i.e. singletons), or their sequence is different by 1bp from a more highly expressed tag (one-offs of highly expressed tags). More specifically, tag sequences that are one-offs are only removed if they do not themselves map to the genome or transcriptome, and if they are expressed at counts below 100. Singleton tags are not significantly different than zero count tags (Audic and Claverie 1997), and are thus not informative in differential expression analyses. If a genome or transcriptome is unavailable, SSOOHE filters cannot be applied,

and instead only singletons and adapters are removed. Other filtering options identify the subsets of tags mapping to genes, and include the "MG" version, containing only tags which map to the genome, the "MR" version, containing only tags which map to the RefSeq transcriptome (Pruitt et al. 2007), and the "MA" version, which contains tags which map to the genome or to any of the Refseq, MGC (Gerhard et al. 2004), Ensembl (Birney et al. 2004), or NCBI predicted transcriptomes (http://www.ncbi.nlm.nih.gov/projects/genome/guide/build.shtml#gene).

2.4.4 Ensembl data

Full gene sequences (including introns), cDNA sequences, and gene boundary coordinates were downloaded from the Ensembl version 47 release, (Birney et al. 2004), based on the NCBI human genome build 36, using the Ensembl API (www.ensembl.org). Virtual sense and antisense tag sequence databases were generated for both full gene and cDNA sequences using in-house Perl scripts. Briefly, all NlaIII sites were identified for each sequence, and the adjoining 17bp in the 3' direction were designated the sense tags, while the 17bp in the 5' direction were designated the antisense tags. The human genome sequence was downloaded from NCBI (ftp://ftp.ncbi.nih.gov), and the complete sequence, including repeat regions, was used to create virtual sense and antisense tag databases. Sense and antisense tag sequences mapping to unique locations in the genome were distinguished from those mapping in multiple locations.

2.4.5 Mouse Tag-seq and LongSAGE replicates

Two libraries from the Mammalian Organogenesis – Regulation by Gene Expression Networks (MORGEN) Project were used to confirm observations made from the human hESC Tag-seq:LongSAGE replicate libraries. The two mouse Tag-seq and LongSAGE replicates were created from heart (Atrioventricular canal) RNA collected at Theiler stage (TS) 17. Libraries were analyzed as described for the human replicate.

2.4.6 Tag-seq vs Affymetrix comparison

Affymetrix exon arrays data was generated at the Genome Sciences Centre. Three sets of Tag-seq:Affymetrix technical replicates were created from acute myeloid leukemia samples (HL60 cell line). Each of the three exon arrays was normalized, yielding a probeset, and a log2plier data file. "Log2plier" is the log of the normalized fluorescence

intensity. Probesets were converted to RefSeq transcripts (Pruitt et al. 2007), and an overall log2plier value was calculated by merging all the probesets for each transcript. Tags in the three Tag-seq replicates were mapped to RefSeq transcripts, and tags mapping to position 1 were retained. A total of 26,000 exon array probesets mapping to RefSeq transcripts were compared to 16,000 position 1 Tag-seq tags. Scatterplots and Pearson correlations were created for all three technical replicate sets.

2.4.7 Tag-seq vs RNA-seq comparison

Blue Stain fungus (*Ophiostoma clavigerum*) RNA was used to create a set of gene expression libraries using the Tag-seq and the RNA-seq protocol. Tag-seq tag sequences were extracted from reads as described above. A total of 2,334,820 reads from the RNA-seq library and 2,334,820 tags from the Tag-seq library were mapped to 11,084 Blue Stain fungus transcripts (DiGuistini et al. 2007). Tag sequences that mapped to the sense strand of annotated transcripts were enumerated separately from those that mapped antisense to known transcripts.

Dynamic range concordance of the two methods was tested by sub-sampling predetermined numbers of tag sequences (in the case of Tag-seq) or reads (RNA-seq), and enumerating the level of expression of each transcript found at those depths (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of the 2,334,820 reads/tags analyzed). For each transcript found in common by both methods (sense-mapping Tag-seq tags only), the Pearson correlation coefficient was calculated between the ten expression values found by Tag-seq vs RNA-seq.

Expression correlations between commonly detected genes were calculated using the sum of all sense and antisense Tag-seq tag counts mapping to each gene, and the sum of the RNA-seq reads mapping to the same gene. RNA-seq values were normalized to gene length (Rosenkranz et al. 2008).

2.4.8 GC-content bias

For this study, we only used tag sequences mapping to the most 3' NlaIII site (i.e. position 1 (P1) tag sequences). The premise was that the GC-content of the observed P1 tag sequences in a library (observed bias) should not be significantly different than the GC-content of a random sample of P1 tag sequences, of the same size, taken from the set

of all possible P1 tag sequences (expected bias). We first measured the deviation of the GC-content of observed P1 tag sequences from that of all possible P1 tag sequences. Second, we randomly sampled the same number of P1 tag sequences from the set of all possible P1 tag sequences (1,000 times), and calculated the deviation of their GC-content from that of the set of all P1 tag sequences. Finally, the deviation of the observed P1 tag sequences was divided by the standard deviation of the deviations of the randomly sampled sets. This value represents the number of standard deviations by which the observed bias in an individual library differed from the expected random bias.

2.4.9 Detection of SAS and isoform ratios between normal and disease samples

Libraries were first grouped by tissue (**Appendix A**), and resulted in a total of 7 tissue groups containing 2 or more LongSAGE libraries (brain, breast, colon, esophagus, gall bladder, retina, and white blood cells); 3 tissue groups contained 2 or more Tag-seq libraries (skin, uterus, and bladder); 5 tissue groups contained both types of libraries (bone marrow, embryonic, lymph nodes, testis, vascular). In each of the 15 tissue library-groups, individual libraries were organized into two subgroups, by whether they represented normal or diseased samples, or different stages of disease. Thus, the libraries in each tissue group could be categorized in multiple ways (**Appendix B**). For each tissue, gene expression between the two subgroups (labeled A and B) was analyzed. Tags in the analysis represented either S-AS gene pairs, single genes with expressed AS tags, or tags representing two isoforms of one gene. Only tags with a minimum expression of 10 tags per million in at least one A or B library were considered. The ratio of the normalized expression (ex. S:AS) was calculated in each A and each B library, and multiplied by the natural log of the difference between them. This ensured that highly expressed tag pairs were more highly ranked. For pairs of genes with positive ratio changes, the S:AS ratio was higher in cancerous tissues versus normal tissues, while pairs with negative values of ratio changes had a S:AS ratio higher in normal rather than cancerous tissues. Gene pair ratio-change values ranged from -11.7 to +11.8.

$$Ratio_{S:AS} = ln\,(\,ln\,(\,exp_S - exp_{AS})\,*\,(\,exp_S / exp_{AS})\,)$$

If the AS was more highly expressed than the sense, the ratio was calculated as:

$$Ratio_{S:AS} = (-1)*(\,ln\,(\,ln\,(\,exp_{AS} - exp_S)\,*\,(exp_{AS} / exp_S)\,)$$

For each tag pair, the average and standard deviation was computed for all modified means in A libraries, and separately for all means in B libraries. An overall measure of the change in S:AS ratio between A and B libraries was therefore:

$$Ratio\ change_{AB} = ln\ (\ mean_A / stdev_A) - ln\ (\ mean_B / stdev_B)$$

Dividing each mean by the standard deviation ensured that tag pairs with lower variance in their ratios were ranked higher than gene pairs with a high variance.

Figure 2.1 Outline of Tag-seq library generation.

Each mRNA (burgundy) underwent double stranded cDNA synthesis using Oligo dT beads, to capture polyadenylated RNA. cDNA (orange) is digested with the NlaIII anchoring restriction enzyme (vertical red arrows), leaving a 4bp overhang (GTAC). Only cDNA fragments anchored to Oligo dT beads are retained. Adapter A (green) is ligated to the overhang, and adds a recognition site for the TypeIIS tagging enzyme MmeI. Following MmeI digestion (red vertical arrow), a second adapter is ligated (Adapter B, blue) to the resulting 2bp overhang. PCR primers (horizontal red arrows) annealing to adapters A and B are used to enrich tags. Cluster generation and sequencing (brown arrow) is performed on the Illumina cluster station and analyzer. The resulting image files are processed to extract the read sequences, and 21bp SAGE tags are further extracted from the reads. Tags consist of the 4bp NlaIII recognition sites and 17bp of unique sequence, and add to 21 bases that can be mapped back to the original mRNA (brown).

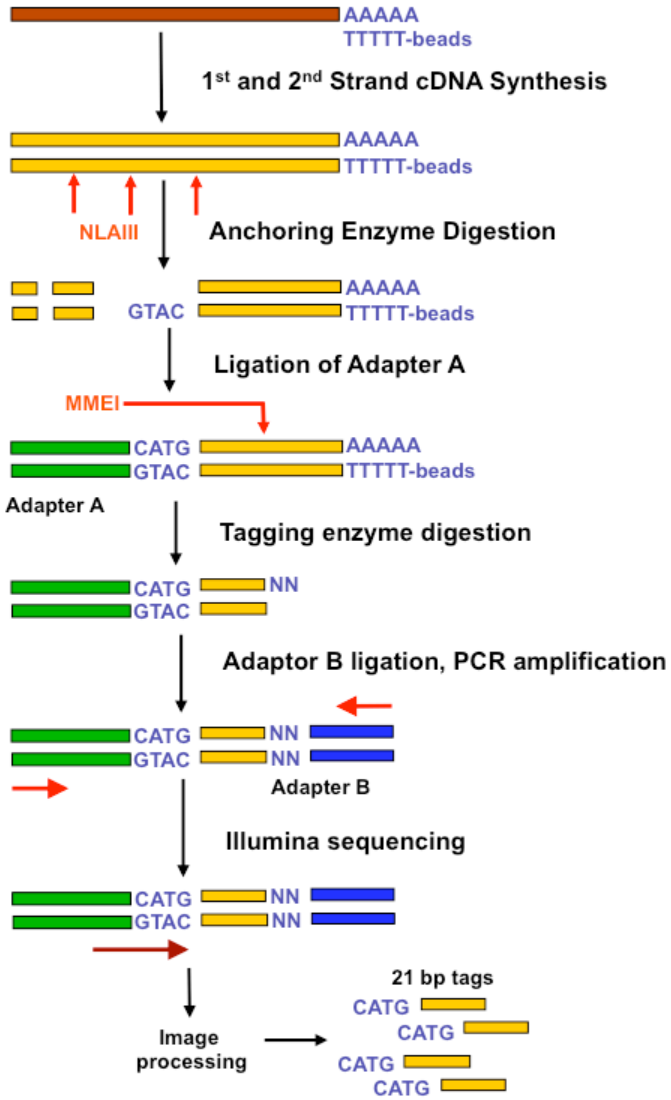
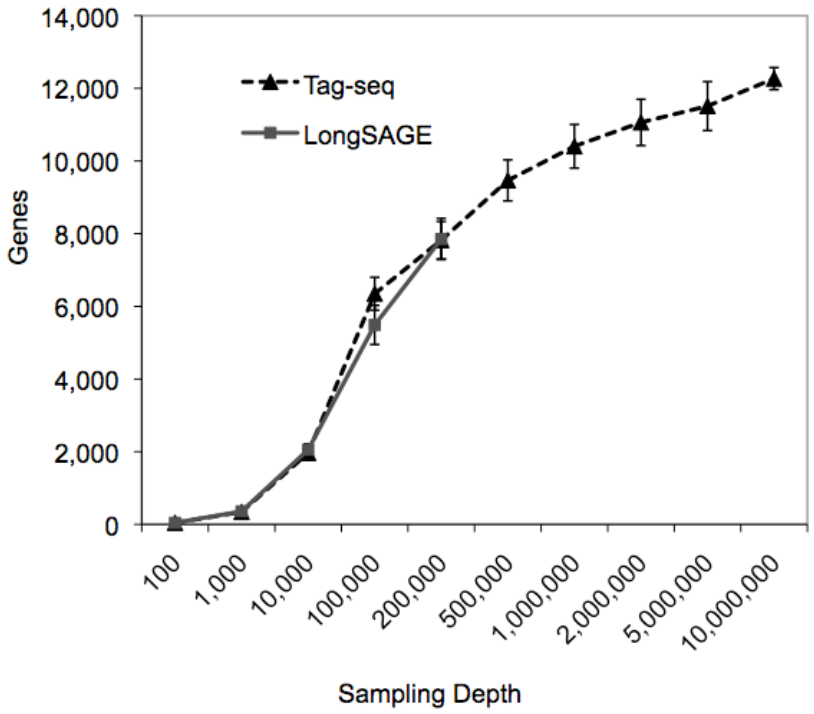


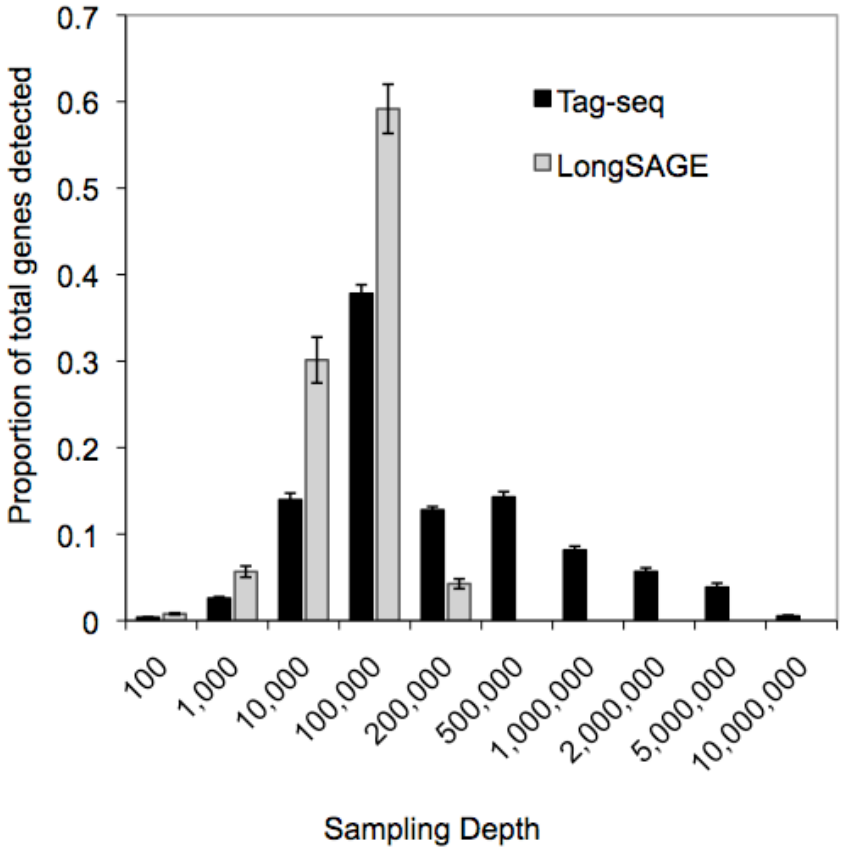
Figure 2.2 Tag to gene mapping success in Tag-seq and LongSAGE.

Average number (A) and proportion (B) of Ensembl genes unambiguously identified in Tag-seq and LongSAGE libraries as a function of sampling depth. Error bars represent the standard deviation of the average number of identified genes in 77 LongSAGE libraries and 35 Tag-seq libraries. The largest LongSAGE libraries were approximately 300,000 tags, while the largest Tag-seq libraries were approximately 10 million tags. (C) Average number of tag sequences found at a minimum count of 10, 20, 50, 100, 200 and 500 in the 35 Tag-seq libraries, which do not map to either the human genome or transcriptome. The number of tag sequences is displayed for Tag-seq only.

A



B


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion of total genes detected
Tag-seq
LongSAGE
100
1,000
10,000
100,000
200,000
500,000
1,000,000
2,000,000
5,000,000
10,000,000
Sampling Depth

C

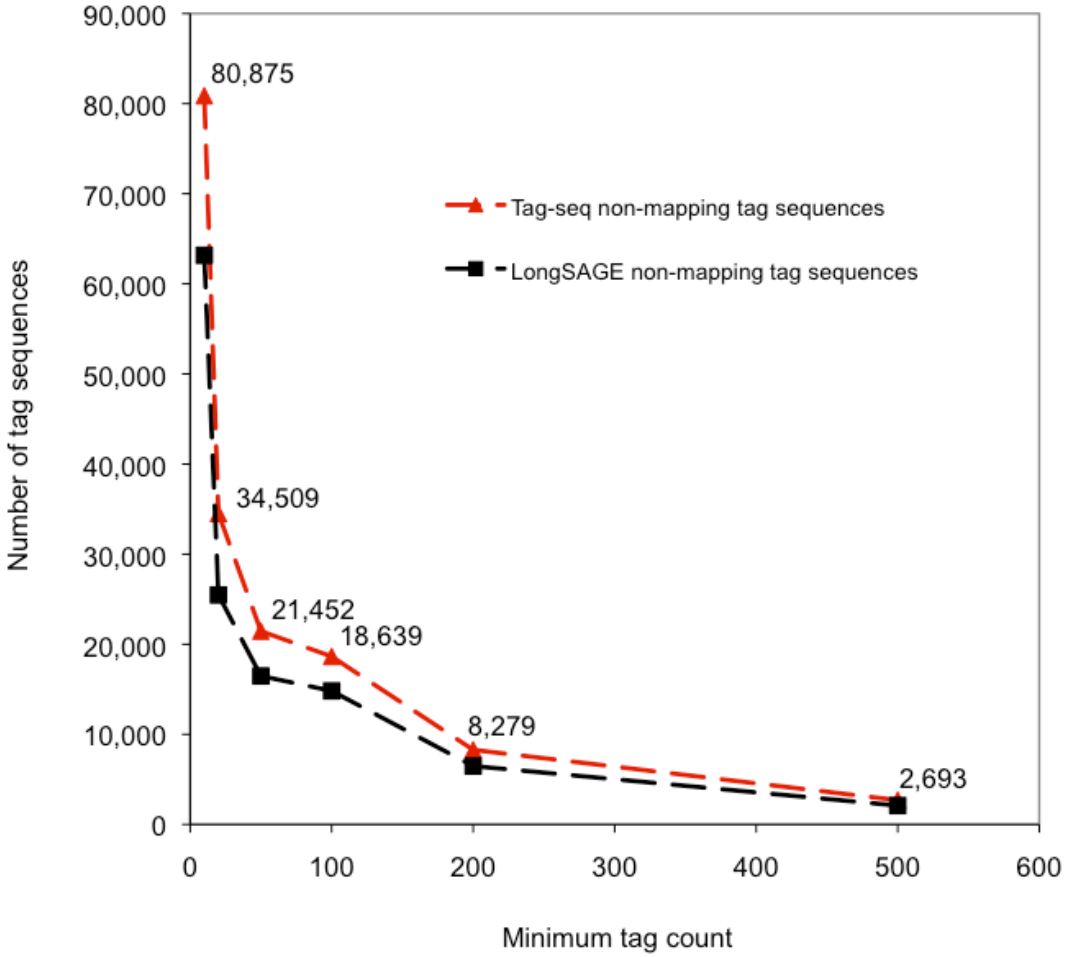
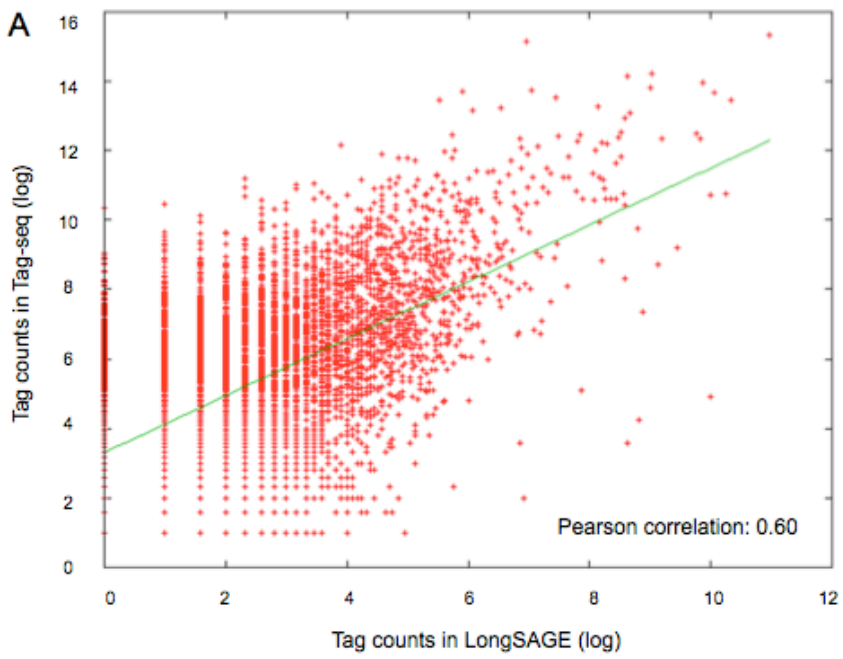
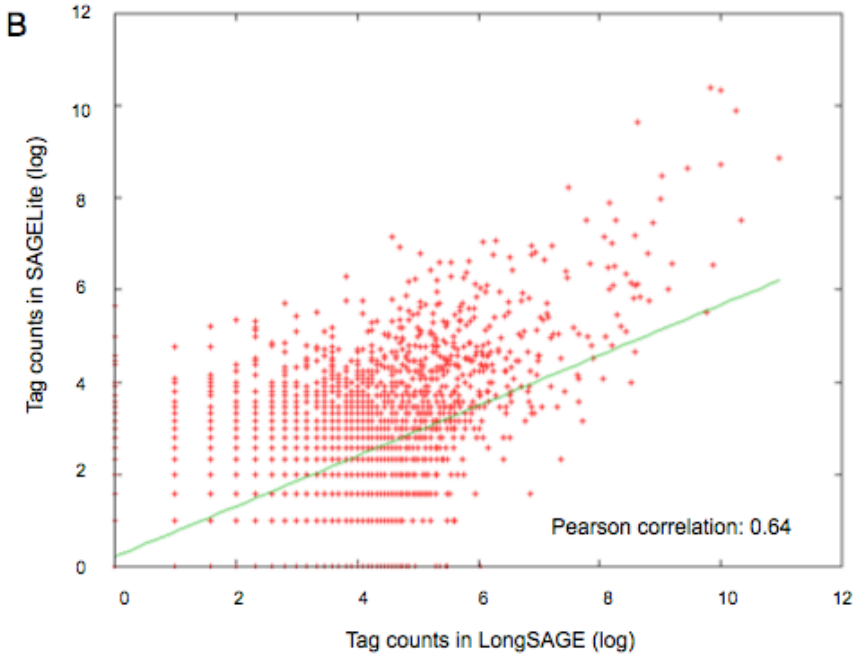
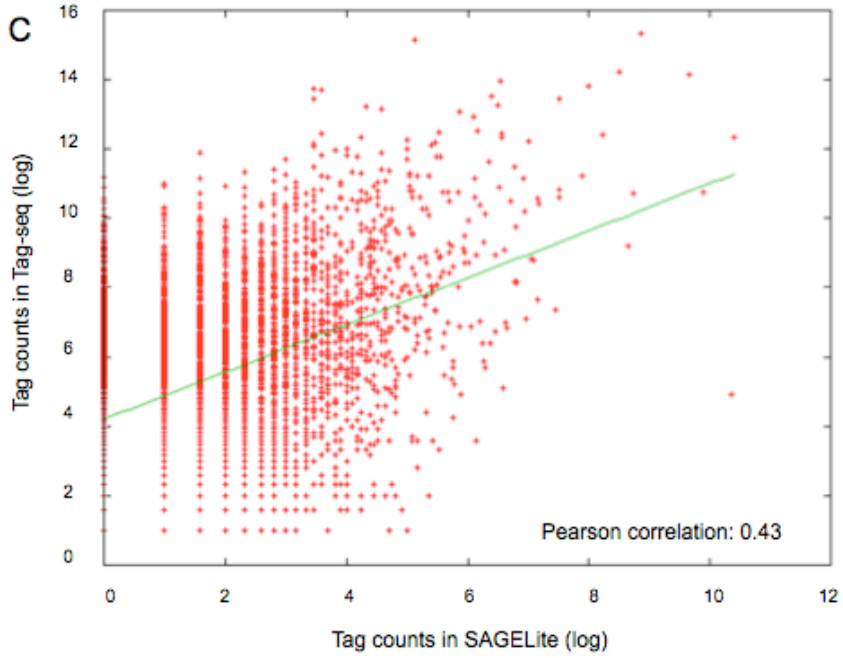

90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Number of tag sequences
0
100
200
300
400
500
600
Minimum tag count
Tag-seq non-mapping tag sequences
LongSAGE non-mapping tag sequences
80,875
34,509
21,452
18,639
8,279
2,693

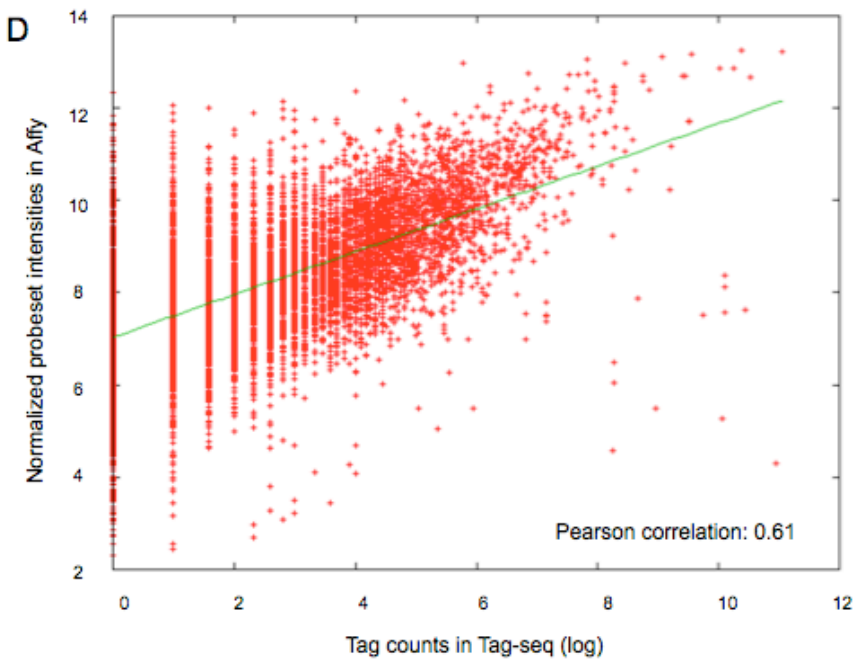
Figure 2.3 Inter-platform comparisons.

Pearson correlations and linear regressions (green lines) are shown for scatterplots of (A) Tag-seq vs LongSAGE replicates, (B) LongSAGE vs SAGELite, (C) Tag-seq vs SAGELite, (D) one of the three technical replicate Tag-seq vs Affymetrix exon arrays. (E) Number of genes detected by three Tag-seq (hs0103, hs0154, hs0155) and three Affymetrix replicate libraries (ex0103, ex0154, ex0155), binned by expression (ie. genes with <10% of max expression in the first bin).


A
Tag counts in Tag-seq (log)
Tag counts in LongSAGE (log)
Pearson correlation: 0.60


B
Tag counts in SAGELite (log)
Tag counts in LongSAGE (log)
Pearson correlation: 0.64


C
Tag counts in Tag-seq (log)
Tag counts in SAGELite (log)
Pearson correlation: 0.43


D
Normalized probeset intensities in Affy
Tag counts in Tag-seq (log)
Pearson correlation: 0.61

E

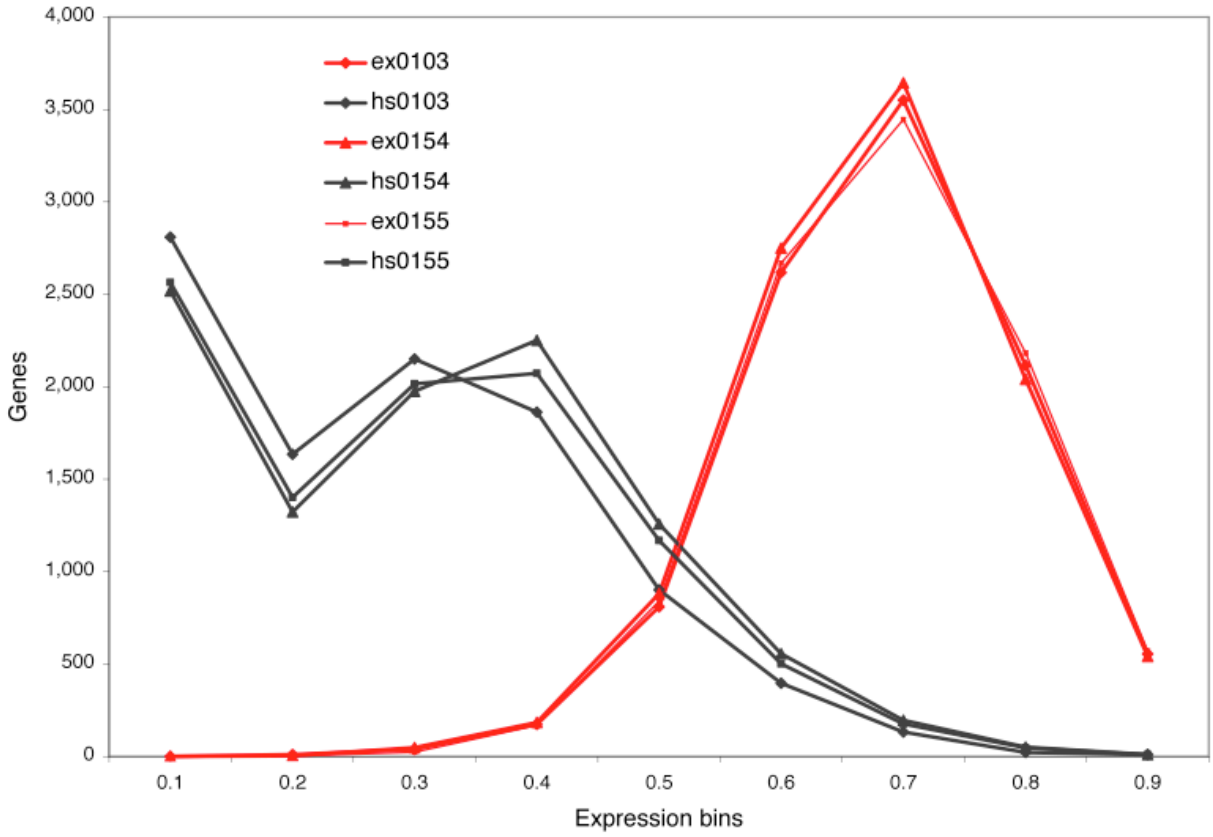
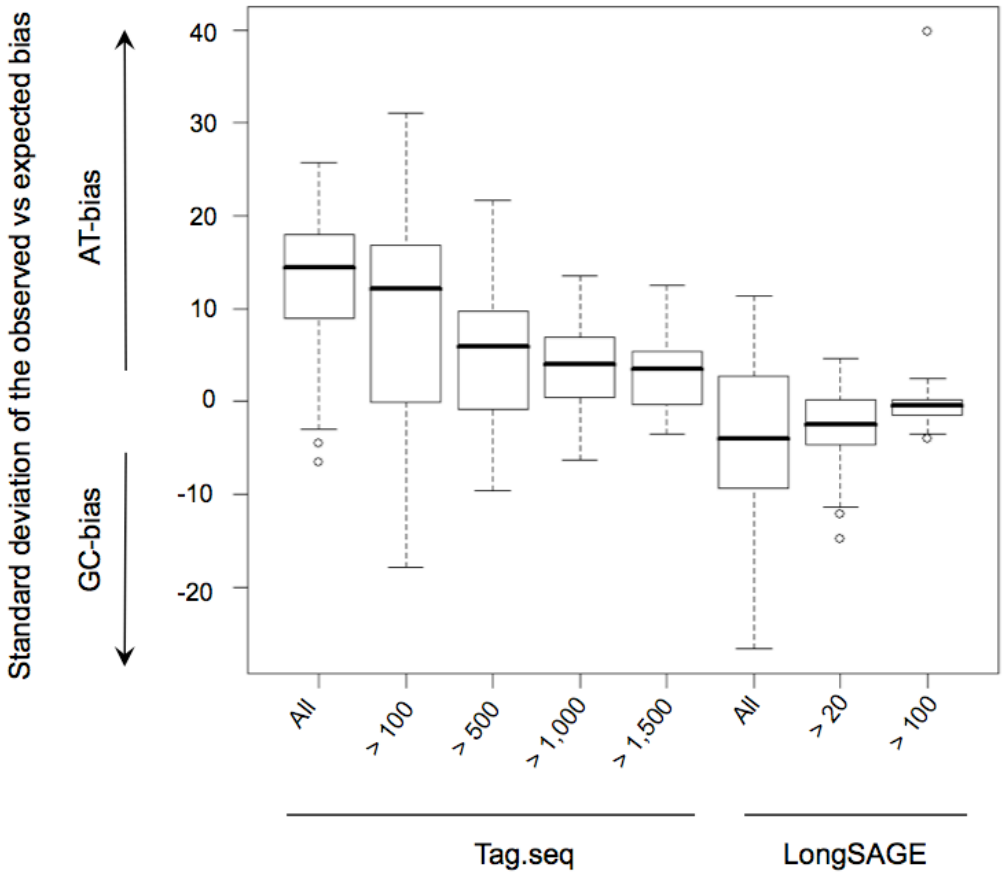

4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Genes
ex0103
hs0103
ex0154
hs0154
ex0155
hs0155
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Expression bins

Figure 2.4 GC-bias of Tag-seq and LongSAGE libraries.

GC-bias is reported as the number of standard deviations by which the observed bias differed from the expected bias (see text). Positive and negative values represent libraries with more AT-rich or CG-rich sequences than expected (respectively). Bias (y-axis) was calculated for all quality filtered Tag-seq and all LongSAGE tag sequences (A) or gene sequences (B), at increasing thresholds of expression (x-axis).

A



B

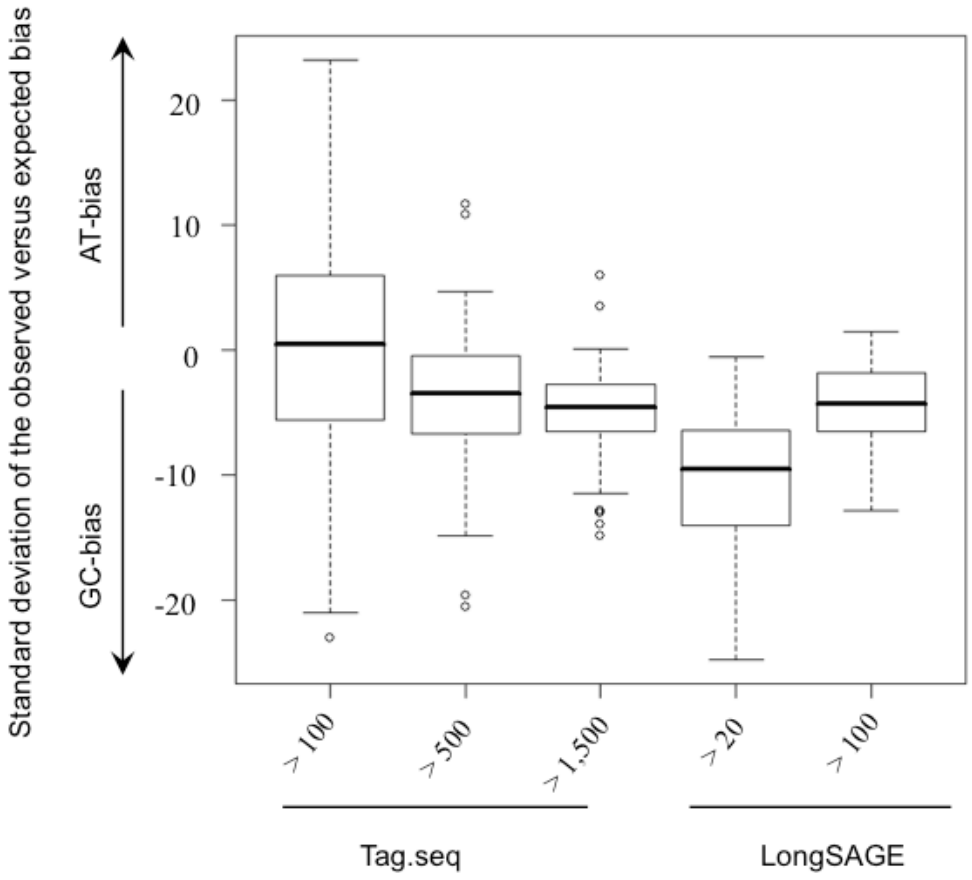
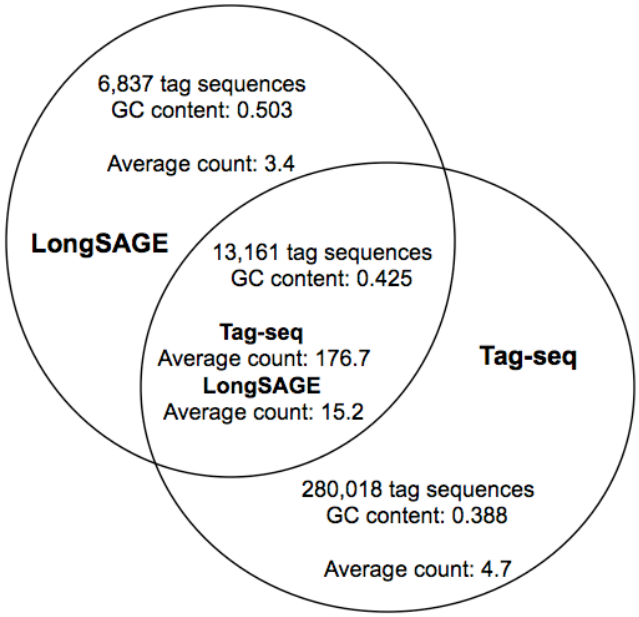

Standard deviation of the observed versus expected bias
AT-bias
GC-bias
20
10
0
-10
-20
> 100
> 500
> 1,500
> 20
> 100
Tag.seq
LongSAGE

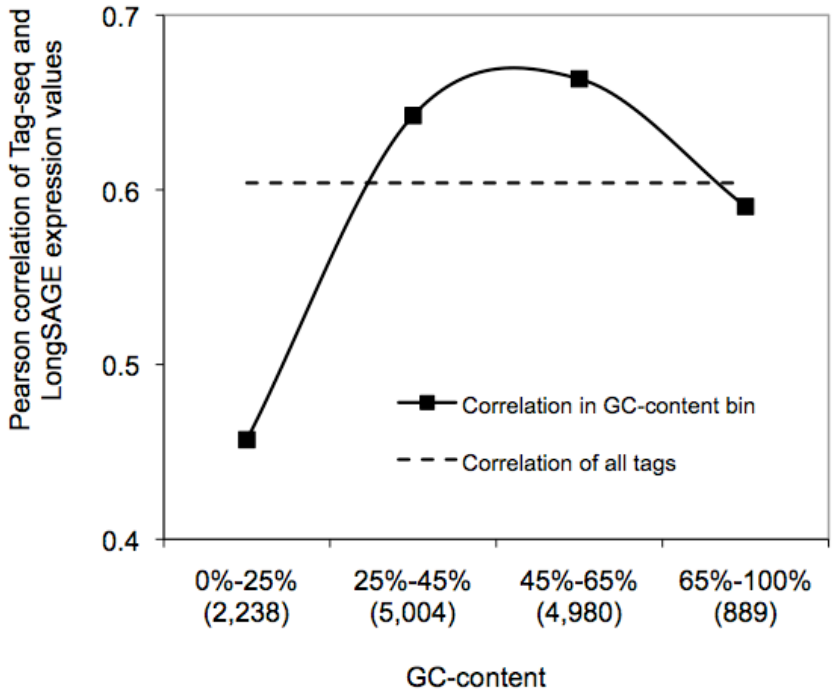
Figure 2.5 GC-content biases in technical replicate libraries.

(A) Comparison of the GC-content and average count of tag sequences found either in common or by each of the Tag-seq and LongSAGE replicate libraries. (B) Pearson correlations of binned tags. Bins are labeled with the range of the observed GC-content, and the number of binned tags (x-axis). (C) Average log expression of tag sequences by bin. An asterix (*) denotes bins between which the expression of tag sequences was significantly different (T-test, $p < 0.01$).

A



B


Pearson correlation of Tag-seq and LongSAGE expression values
0.7
0.6
0.5
0.4
Correlation in GC-content bin
Correlation of all tags
0%-25% (2,238)
25%-45% (5,004)
45%-65% (4,980)
65%-100% (889)
GC-content

C

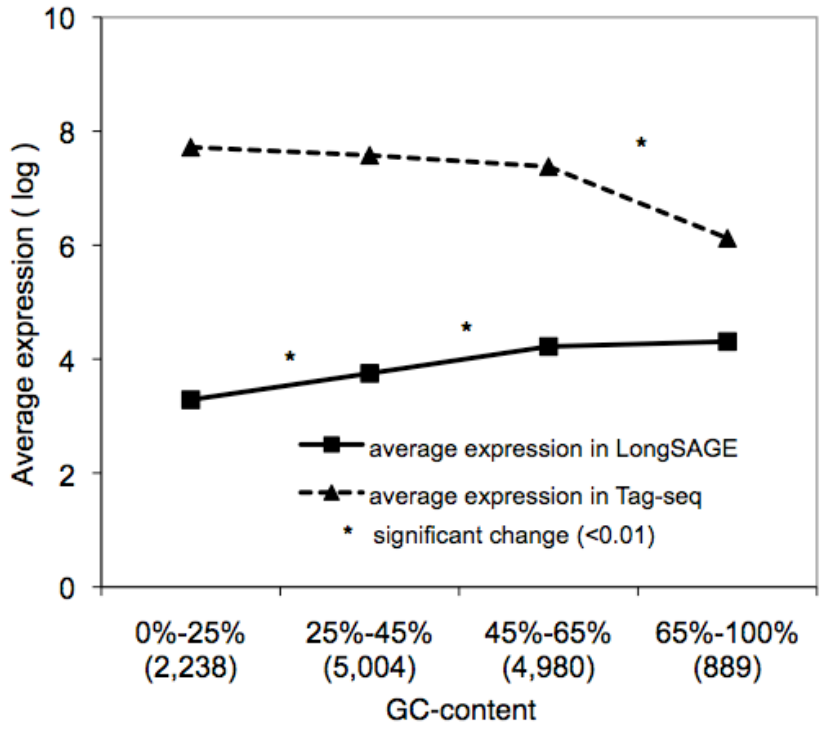

Average expression (log)
10
8
6
4
2
0
average expression in LongSAGE
average expression in Tag-seq
* significant change (<0.01)
0%-25% (2,238)
25%-45% (5,004)
45%-65% (4,980)
65%-100% (889)
GC-content

Figure 2.6 Sense and antisense trags of cis-encoded antisense genes and non-overlapping bi-directional genes.

(A) Diagram of a cis-encoded S-AS locus with two genes overlapping in a convergent manner. (i) Exons are open rectangles, and introns are horizontal lines connecting exons. Arrows denote transcriptional start sites. NlaIII restriction sites (vertical lines) are numbered starting with the 3' most site, and the virtual sense (green) and antisense (red) tags at each position are shown on both strands. (ii) Experimentally observed tags can map in a sense orientation to exons or introns of the gene they arise from (green rectangle). The sense tags originating from the #1 position of the top-strand gene also maps antisense at the #2 position of the bottom-strand gene. The sense tag originating from the #3 position on the top-strand gene is outside the bottom-strand gene boundary, so there is no corresponding antisense tag in the bottom-strand gene. (iii) cDNA sequences exclude intronic tags. (B) Antisense tags that map to a location with no annotated transcript on the opposite strand are evidence of novel transcription. (i) Virtual and (ii) experimentally observed tags are diagramed. The tag mapping antisense at the #1 position of the bottom-strand gene was generated from an unannotated region of unknown structure (grey box) on the top-strand. (iii) The sum of the sense cDNA tags of annotated genes can be compared to the sum of all tags mapping on the opposite strand but within the gene boundaries of the annotated gene.

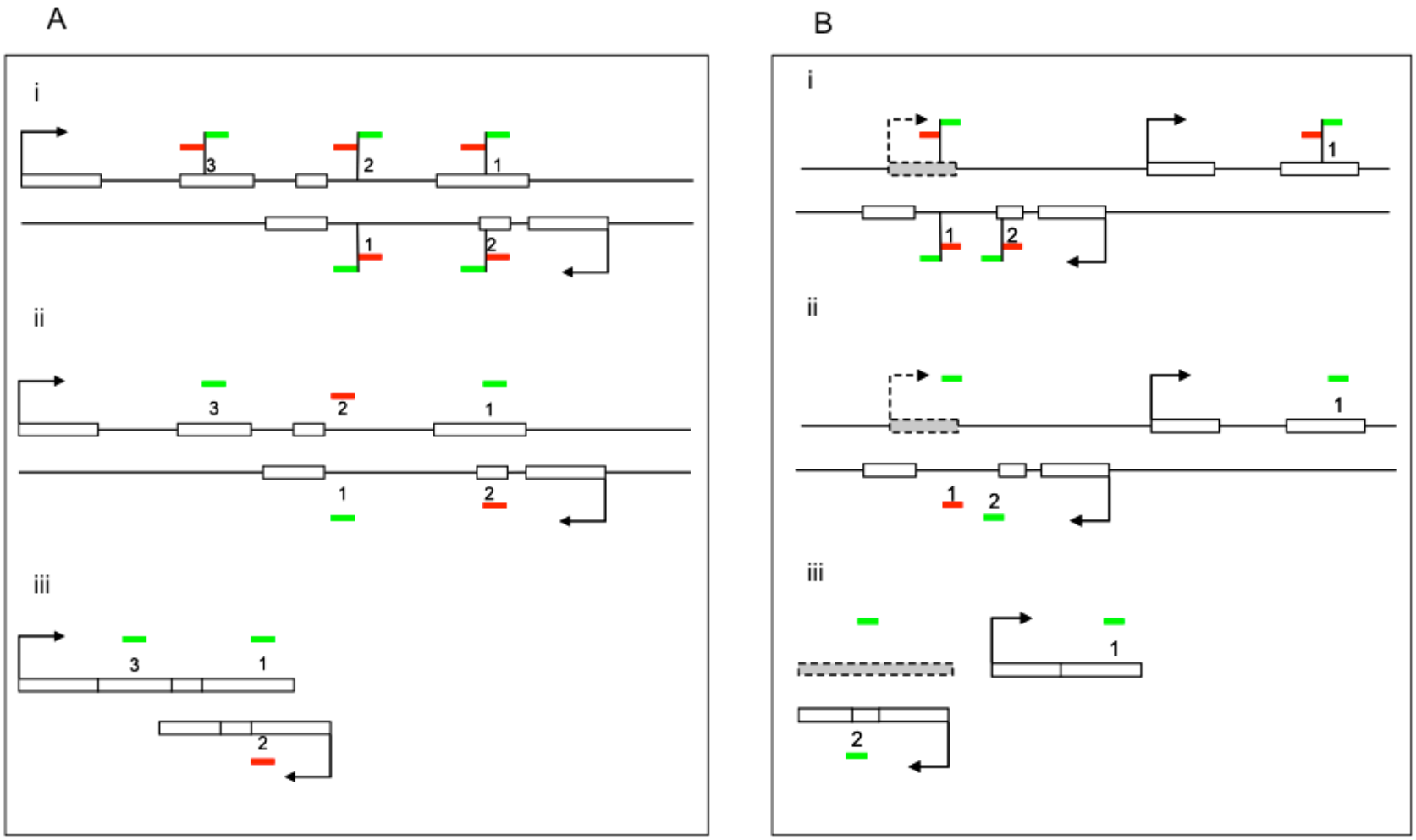
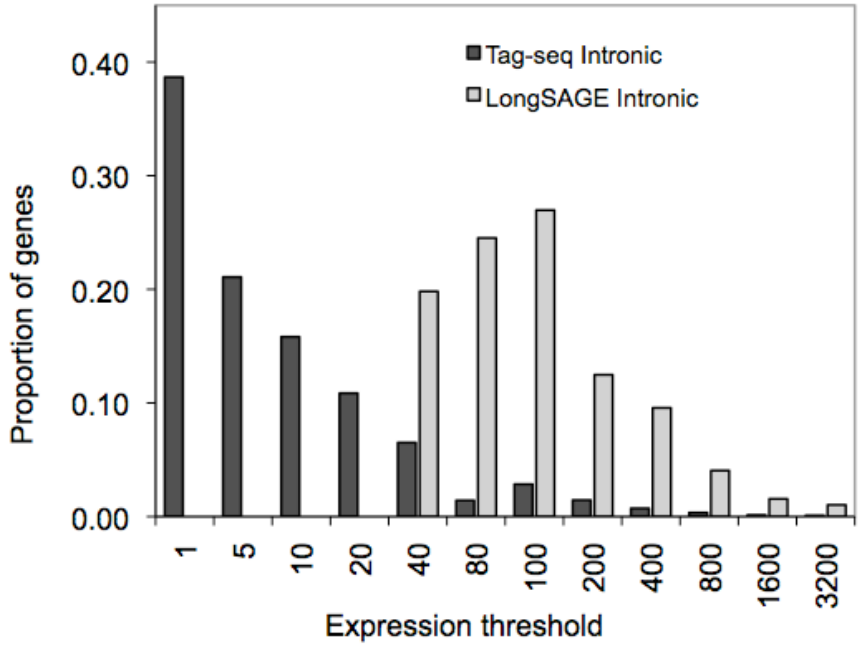

A
i
3
2
1
1
2
ii
3
2
1
1
2
iii
3
1
2
B
i
1
1
2
ii
1
1
2
iii
1
2

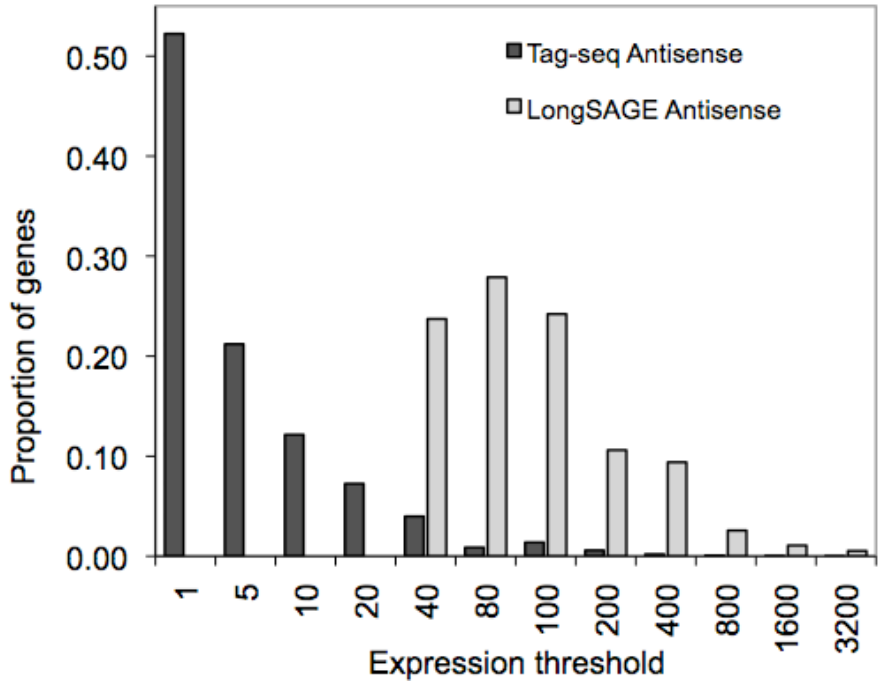
Figure 2.7 Detection of intronic, antisense, and TFs by Tag-seq and LongSAGE.

The proportion of the average number of genes detected by tags in LongSAGE and Tag-seq libraries is shown at increasing expression thresholds (tags per million). Bars represent the proportion of the average number of genes with (A) intronic tags, (B) antisense tags, and (C) DNA-binding domains (transcription factors) in Tag-seq and LongSAGE libraries.

A



B


Tag-seq Antisense
LongSAGE Antisense
Proportion of genes
0.50
0.40
0.30
0.20
0.10
0.00
1
5
10
20
40
80
100
200
400
800
1600
3200
Expression threshold

C

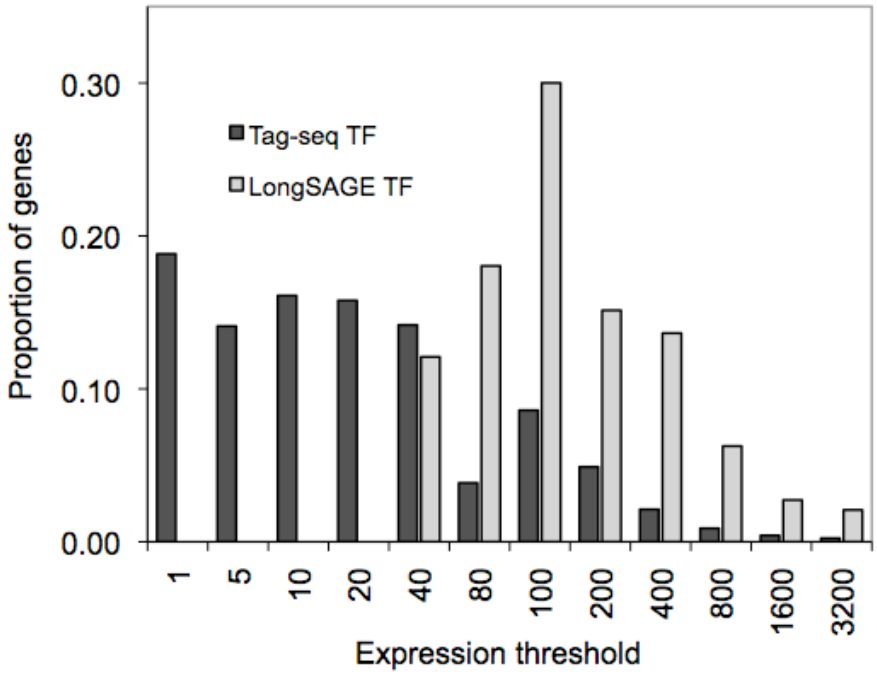
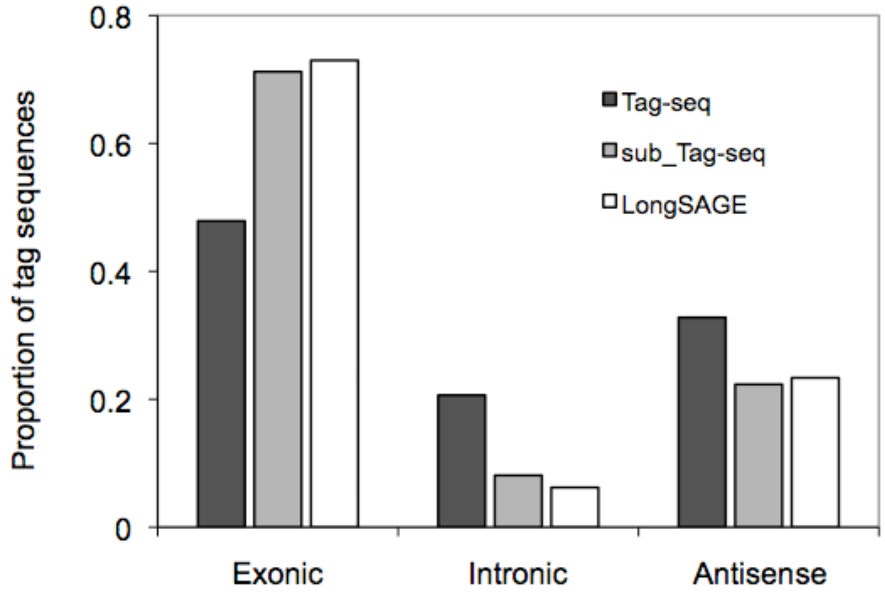

Tag-seq TF
LongSAGE TF
Proportion of genes
0.30
0.20
0.10
0.00
1
5
10
20
40
80
100
200
400
800
1600
3200
Expression threshold

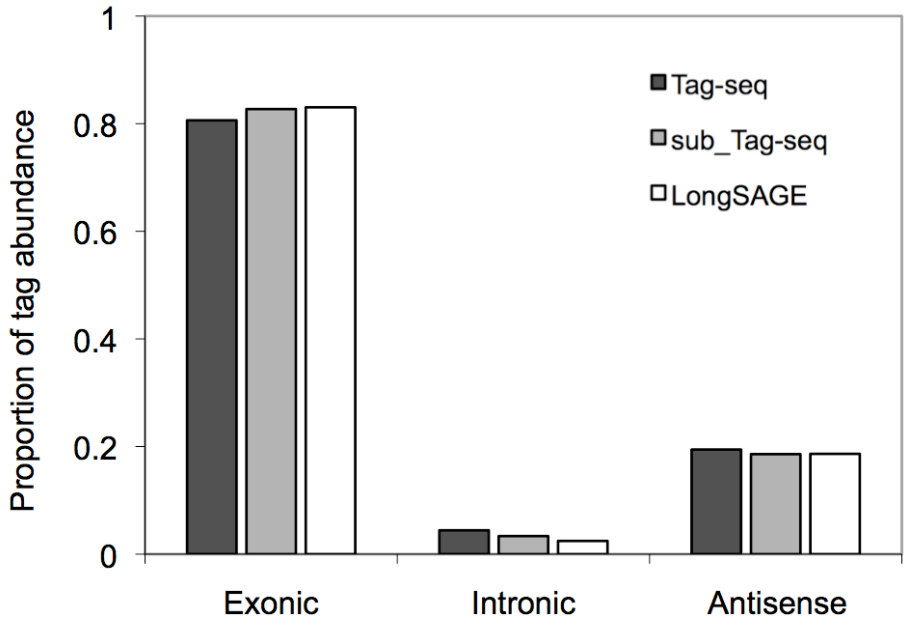
Figure 2.8 Detection of intronic, antisense, and TFs in replicate libraries.

Tag sequences from the hESC Tag-seq technical replicate, the in silico derived sub_Tag-seq, and the LongSAGE replicate, were mapped to the introns, exons, and antisense strands of Ensembl genes. The proportions of distinct tag sequences (A) and tag abundance (B) are reported relative to all mapped quality-filtered tags. Average tag counts (+/- standard deviation) are reported for all tag sequences found in common between the three libraries (C).

A



B


Proportion of tag abundance
1
0.8
0.6
0.4
0.2
0
Tag-seq
sub_Tag-seq
LongSAGE
Exonic
Intronic
Antisense

C

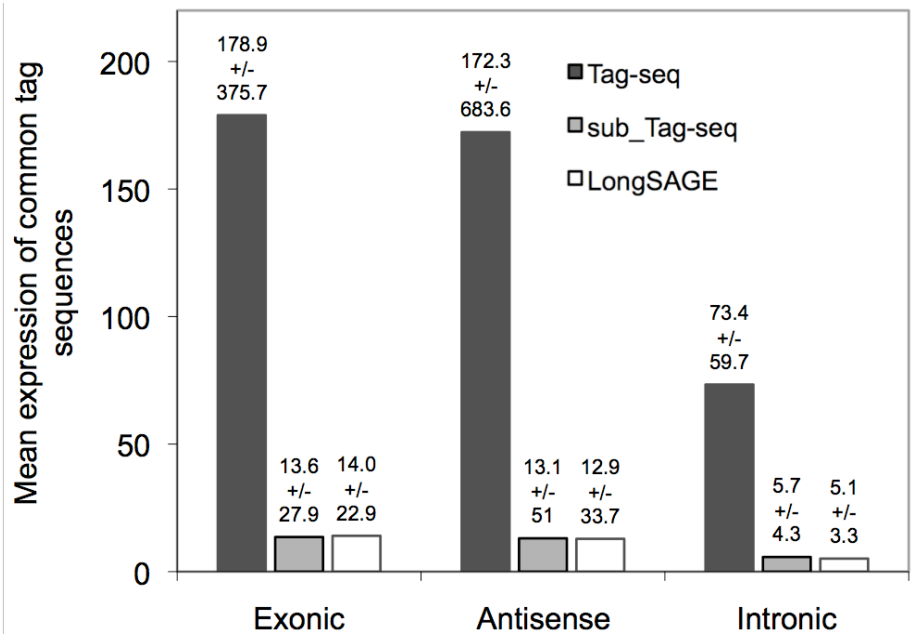
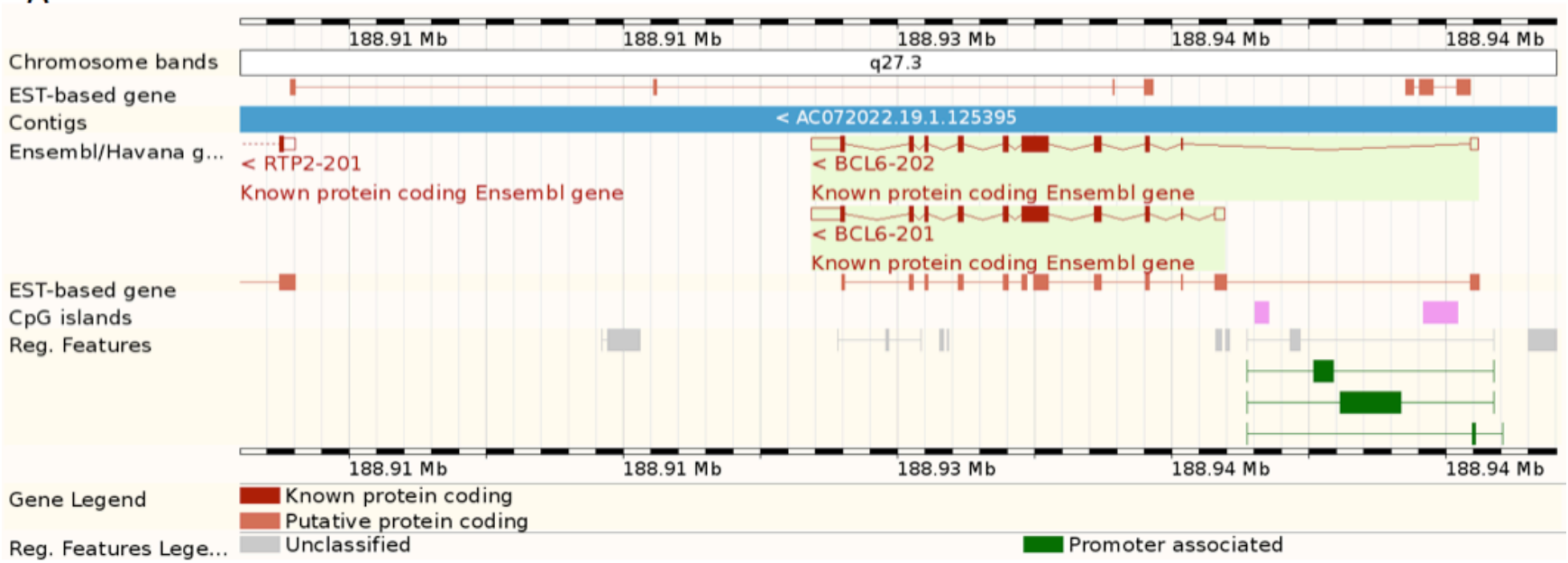

Mean expression of common tag sequences
200
150
100
50
0
178.9 +/- 375.7
13.6 +/- 27.9
14.0 +/- 22.9
172.3 +/- 683.6
13.1 +/- 51
12.9 +/- 33.7
73.4 +/- 59.7
5.7 +/- 4.3
5.1 +/- 3.3
Tag-seq
sub_Tag-seq
LongSAGE
Exonic
Antisense
Intronic

Figure 2.9 Change in the ratio of BCL6 sense and antisense tags in breast cancer.

(A) The novel SAS locus, BCL6, has antisense ESTs and antisense tags. (B) Antisense transcription is significantly more abundant in libraries representing grade II carcinomas (class A) compared to remaining tissues (class B). Observed tag expression (counts per million) is shown for breast cancer patient IDC7.

A



B

Library	BCL6S	BCL6 AS	Classification
644 breast - Carcinoma associated myofibroblast (grade II)	107.55	76.82	A
645 breast - Carcinoma epithelium (gradeII)	108.97	13.62	A
646 breast - Carcinoma associated stroma	73.5	0	A
647 breast - Normal myoepithelium	72.46	0	B
659 white blood cells - Breast carcinoma	29.07	0	B

Table 2.1 Tag sequences detected by Tag-seq and LongSAGE

Average expression values were enumerated for tag sequences detected only in LongSAGE libraries, only in Tag-seq libraries, or in both. The average number of libraries in which individual tags are expressed is also shown. Results are tabulated for all tags or the subset of tags mapping to Ensembl genes. SD = standard deviation.

		Common Tags		
	LongSAGE	**LongSAGE**	**Tag-seq**	**Tag-seq**
Tag sequences (all)	822,988	318,400	318,400	3,705,783
Average number of libraries +/- SD	1.3 +/- 1.3	4.9 +/- 9.6	7.4 +/-9.1	1.7 +/- 1.8
Average expression level +/- SD	1.1 +/- 1.8	3.7 +/- 20.9	62.1 +/-1115.2	3.8 +/- 71.9
Tag sequences (Ensembl genes)	543	98,717	98,717	1,026
Ensembl genes detected	432	21,638	21,638	741
Average number of libraries +/- SD	1.6 +/- 1.5	5.1 +/- 10.5	5.7 +/- 8.5	2.9 +/- 4.0
Average expression level +/- SD	2.5 +/- 7.9	3.9 +/- 21.9	65.8 +/- 1116.0	73.8 +/- 584.0

Table 2.2 Known and novel SAS genes in the Cancer Gene Census

(A) The proportion of known and novel SAS genes in the cancer gene census with altered sense to antisense expression ratios (AR). Genes are sub-categorized into those with expression ratio scores in the top 20% and top 10%. (B) The proportion of DE SAS genes in the cancer gene census, of 300 observed in CGAP.

A

	Genes	Cancer Census	Proportion
Known SAS	389	6	0.015
Top 20% AR	264	5	0.019
Top 10% AR	193	3	0.016
Novel SAS	2,195	72	0.033
Top 20% AR	1,337	39	0.029
Top 10% AR	935	27	0.029

B

	Cancer census	Proportion (of 300)
Altered SAS ratios	78	0.26
Top 20%	44	0.56
Top 10%	30	0.38

Table 2.3 miRNA targeting sites

Genes with differentially expressed (DE) isoforms between two disease states were enriched in miRNA targeting sites relative to all genes. The enrichment was further increased for genes with DE values in the top 20% and most increased for those in the top 10%.

	Genes	Genes with miRNA targeting sites	Proportion
All Ensembl genes	33,761	7,442	0.22
Genes with DE isoforms	1,957	1,304	0.67
Top 20% DE	1,156	806	0.7
Top 10% DE	772	560	0.73

Table 2.4 Affymetrix versus Tag-seq comparisons

(A) Tag sequences detected in common by the Affymetrix and the Tag-seq platforms were analyzed for three sets of replicate libraries. Minimum, mean, and maximum expression (log2) was calculated for the tag sequences in each library. The range between the minimum and the maximum, and between the mean and maximum values was computed. (B) Genes detected by either the RNA-seq and the Tag-seq platforms, as well as in common between the platforms, were analyzed for one set of replicate libraries. A minimum of two RNA-seq reads and two Tag-seq tags were required to count a gene as expressed. The total number of genes with sense tags (S), with antisense tags (AS), or with both sense and antisense tags (S-AS) are enumerated.

A

	Replicate 154		Replicate 155		Replicate 103	
	Affy	**Tag-seq**	**Affy**	**Tag-seq**	**Affy**	**Tag-seq**
Tag sequences in common	10,152	10,152	9,969	9,969	9,930	9,930
Mean expression (log2)	8.35	2.84	8.41	2.73	8.45	2.51
Min expression (log2)	2.31	0.00	2.30	0.00	2.00	0.00
Max expression (log2)	13.25	11.04	13.24	10.87	13.33	10.89
Range (Max-Min)	10.94	11.04	10.94	10.87	11.32	10.89
Mean Range (Max-Mean)	4.90	8.20	4.82	8.14	4.88	8.38

B

	Total genes	**Genes in common**
RNA-seq (all)	8,528	8,050
Tag-seq (all)	8,366	8,050
Tag-seq (S)	5,224	5,064
Tag-seq (AS)	3,142	2,986
Tag-seq (S-AS)	2,430	2,373

3. Extensive relationship between antisense transcription and alternative splicing in the human genome[4]

Author contributions

A.S.M. and M.A.M. conceived the analyses. A.S.M. designed and performed all computational analyses, created the figures and tables, and wrote the manuscript. M.G. conducted microarray data pre-processing (section 3.4.3.1) and contributed microarray analysis concepts.

3.1 Introduction

Much of the complexity of mammalian biology can be attributed to the regulation of gene expression via changes in the level, splicing, and localization of RNA (Wang et al. 2008a; Licatalosi and Darnell 2010). One type of regulation occurs between genes that are encoded in an overlapping and opposite orientation. Such sense-antisense (SAS) gene pairs encode proteins and non-coding RNAs that play key roles in development, and have been implicated in diseases such as cancer (Vanhee-Brossollet and Vaquero 1998; Tufarelli et al. 2003; Reis et al. 2004; Chen et al. 2005; Engstrom et al. 2006). Antisense transcripts have been identified at 50% to 70% of mammalian loci (Riken Genome Exploration Research et al. 2005), yet despite their prevalence, regulatory roles have only been elucidated for a small subset of SAS genes (reviewed in (Vanhee-Brossollet and Vaquero 1998; Lavorgna et al. 2004). Since a large proportion (40%) of antisense transcripts are non-coding RNAs, they may act predominantly as regulators of expression (Mattick 2004).

In a limited number of cases, antisense transcription has been correlated to sense gene splicing (Mihalich et al. 2003; Louro et al. 2007; Annilo et al. 2009), or shown to regulate sense gene splicing (Krystal et al. 1990; Kuersten and Goodwin 2003; Yan et al. 2005; Beltran et al. 2008). One well-characterized example is the antisense-mediated splicing regulation of the thyroid hormone receptor (*TR*α) by the antisense transcript *Rev-erbA*-α (Hastings et al. 1997). At this locus, co-expressed sense and antisense transcripts

[4] A version of this chapter has been submitted. Morrissy AS, Griffith M, Marra MA. 2010. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Submitted.*

form double-stranded RNA (dsRNA) over the region of SAS overlap, leading to splice site masking and a consequent shift in mRNA isoform production. Similar changes in splicing can be achieved by the addition of synthetic antisense oligonucleotides (Garcia-Blanco et al. 2004). In vitro and in vivo, synthetic antisense oligonucleotides have been used to modulate splicing reactions in favor of specific isoforms of disease-related genes, with the goal of developing strategies that influence therapeutic outcomes (Garcia-Blanco et al. 2004).

To date, there are no genome-wide studies that have investigated the relationship between alternative splicing and antisense transcription. We therefore set out to investigate the possibility that antisense-mediated splicing may be a prevalent regulatory mechanism in the human genome. The specific objectives of our study were to assess the correlation between antisense transcription and splicing events in normal human cells, and to investigate possible mechanisms for antisense-mediated splicing regulation.

3.2 Results

3.2.1 Distinct structural features of known novel SAS loci

The hypothesis that antisense transcription has the capacity to influence splicing outcomes implies that genes with antisense transcripts may be structurally distinct from those without. We therefore compared structural features (gene length, transcript length, number of introns, and number of annotated isoforms) of genes with and without antisense transcripts. Genes with antisense transcripts were classified into two categories: (1) those with an annotated antisense gene whose genomic coordinates overlapped ("known SAS"), and (2) those with no annotated antisense gene, but with evidence for antisense transcription ("novel SAS") (**Fig. 3.1A**). We analyzed available Ensembl annotations for 20,921 protein-coding genes, and identified 5,169 known SAS genes (**Fig. 3.1A**). To find evidence for novel antisense transcription, we searched for Affymetrix Human Exon 1.0 ST array probesets mapping on the antisense strand of annotated genes without partners (Methods, **Fig. 3.1B**). Probesets were used to infer evidence for novel antisense transcription at 7,823 genes (reffered to as "novel SAS genes"). An additional 7,929 genes had no evidence for either known or novel antisense transcription ("non SAS").

On average, known SAS and novel SAS genes were significantly longer than non SAS genes (83.7 kb and 70.6 kb, versus 22.9kb, $P < 1.6\times10^{-148}$), their transcripts were larger (2.3 kb and 2.6 kb versus 1.9 kb, $P < 5.1\times10^{-40}$), they had more introns (9.4 and 9.9 versus 6.6, $P < 5.8\times10^{-78}$), and more isoforms (2.3 and 2.3, versus 1.8, $P < 3.2\times10^{-84}$) (**Fig. 3.1C**, all Welch t-Test P values in **Fig. 3.1D**). The co-incidence of antisense transcription with longer genes may reflect an increased chance of observing antisense transcription in larger genomic regions, similarly to the increased numbers of exons and isoforms observed at such loci. However, given previous observations of antisense-regulated splicing events (Krystal et al. 1990; Hastings et al. 1997; Kuersten and Goodwin 2003; Yan et al. 2005; Beltran et al. 2008), these results are consistent with a putative role for antisense transcription in splicing regulation.

3.2.2 Functional characterization

Structurally and functionally, novel SAS genes were most similar to known SAS genes, and most dissimilar to genes without antisense transcripts (**Appendix E**). The novel SAS gene category was the most enriched in functional categories, including 58 GO categories and 45 Keywords. Notably, 31 of 39 GO terms in the Biological Process category pertained to regulation (i.e. “Regulation of Apoptosis”, “Negative Regulation of Transcription”, “Regulation of Signal Transduction”, etc). Highly enriched UniProt Keywords included “Phosphoprotein”, “Alternative Splicing”, “Kinase”, “Apoptosis”, and “Proto-oncogene”. Overall, these GO terms and Keywords are consistent with the strong enrichment of Cancer Gene Census (CGC) genes observed in the novel SAS class relative to the set of all protein coding genes (215 of 389 CGC genes, Chi-square test, $P = 4.2\times10^{-9}$).

3.2.3 Exon splicing is strongly correlated to antisense gene expression

Having established that antisense transcription is generally associated with structurally distinct genes having multiple isoforms, we examined the relationship between alternative splicing and antisense transcription at these loci. To establish the parameters of this relationship in normal human tissues, we analyzed expression data derived from 176 lymphoblastoid cell lines (LCL’s) (Huang et al. 2007). These data were generated using the Affymetrix Human Exon 1.0 ST arrays which measure expression changes for 1.4 million probesets representing known and predicted exons on both strands of the

genome. Eighty-seven Centre d'Etude Polymorphisme Humain individuals from UTAH (CEU) and 89 Yoruba individuals from Iabadan, Nigeria (YRI) were included in the analysis.

Probesets mapping to the sense strand of Ensembl exons were used to measure sense gene expression (**Fig. 3.1A**). Similarily, probesets mapping to opposite strands of genes without an annotated antisense gene partner measured novel antisense transcription (**Fig. 3.1B**) since they were designed based on previous evidence of transcription, such as ESTs (expressed sequence tags) (Liu et al. 2003). Probesets mapping to introns were also analyzed since these may represent alternative splice variants of annotated genes, including those with novel exons, or exons with alternative 3′ and 5′ splice sites. Therefore, each analyzed probeset mapping to a gene was considered in our analysis to represent an exon.

The alternative splicing of each exon was assessed by normalizing its expression value to the expression value of the gene in each sample (as described in Methods). The resulting value (denoted the splice index) represented an exon's relative inclusion or exclusion from the final mRNA. Thus, correlating the splicing index of individual exons with the expression of the antisense gene allowed us to determine the relationship between splicing and antisense transcription (Methods). Overall, a total of 2,995 exons in 258 known SAS genes were expressed in the LCL samples, as well as 4,187 exons in 215 novel SAS genes. P-values were corrected for multiple-testing using the stringent Bonferroni method, and we therefore expect our results to be conservative.

Our analysis revealed a widespread relationship between splicing and antisense transcription in human LCL's. Of the 258 known SAS genes, a large majority (191 genes, 74.1%) had antisense-correlated splicing events (Methods). Overall, the splicing index of 24% of the 2,995 expressed exons in these genes was significantly correlated to antisense gene expression (**Fig. 3.3**, Bonferroni corrected $P < 0.05$). Of these 191 known SAS genes, 75.4% had antisense-correlated splicing changes in both partners, as would be expected from a reciprocal relationship. An example of this reciprocal relationship is the MSH6 (mutS homolog 6) and FBXO11 (F-box protein 11) locus, which encodes 9 exons with significant antisense-correlated splicing (**Fig 3.2A**). The two MSH6 exons were profiled by three distinct probesets. Two of these had splice index values negatively

correlated to the antisense gene (FBXO11, **Fig. 3.2B**), and one had splice index values positively correlated to the antisense gene (**Fig. 3.2C**). This indicated that the negatively correlated probesets were excluded from MSH6 mRNA isoforms expressed concurrently with FBXO11, while the positively correlated exon was preferentially included. Interestingly, the MSH6 exon that encodes the protein motif for DNA mismatch repair was profiled by two probesets (**Fig 3.2A**). The 5′-most probeset had a positively antisense-correlated splice index (r = 0.56), and the 3′-most probeset has a negatively antisense-correlated splice index (r = -0.59). The 3′-most of these probesets maps within the DNA mismatch repair motif (data not shown), and thus distinguishes those MSH6 isoforms that contain the motif from shorter isoforms that do not. Since the splicing index of this probeset (as well as that of another downstream probeset with r = -0.63, **Fig. 3.2A-B**) is negatively correlated to antisense expression, it seems that FBXO11 expression is positively correlated to short and presumably non-functional MSH6 isoforms.

Similarly to the known SAS genes, 78.1% (168) of the 215 novel SAS genes had significant antisense-correlated splicing events. A total of 19.8% of the 4,167 expressed exons in these genes had antisense-correlated splicing patterns (**Fig. 3.3**). Genes contained (1) exons with positive antisense-correlated splicing, indicating their inclusion in isoforms co-expressed with the antisense gene; (2) exons with negative antisense-correlated splicing, indicating their exclusion from expressed isoforms; and (3) exons whose splicing was un-correlated with antisense transcription, indicating either constitutive expression or splicing regulation mediated by independent factors. On average, 32.3% of known and 23.1% of novel SAS gene exons had antisense-correlated splicing, suggesting that expressed alternative isoforms differed significantly from each other. Over a third of exons with antisense-correlated splicing events encoded protein domains (data not shown), and may therefore alter the encoded protein.

We re-analyzed the CEU and YRI data separately, since previous studies (Spielman et al. 2007; Storey et al. 2007; Zhang et al. 2008) have observed population-specific differences in gene expression patterns. Such differences were also evident in our data, as a remarkably higher proportion of novel SAS genes undergo antisense-correlated splicing solely in the YRI (35.6%) versus CEU (24.1%) individuals (**Fig. 3.4**).

3.2.4 Antisense expression affects both splicing and expression of sense genes

Previous studies have identified correlations between antisense transcription and sense gene expression (Chen et al. 2005; Riken Genome Exploration Research et al. 2005), however, the correspondence between antisense-correlated changes in splicing versus gene expression remains to be determined. Using the 258 known SAS genes expressed in LCLs, we calculated correlations between gene expression levels of partner genes and found significant gene-level correlations for 68.2% of pairs (176 genes, Bonferroni corrected $P < 0.05$). Given that antisense-correlated splicing occurs at 74.0% of the same 258 known SAS genes, these results indicate that antisense transcription affects splicing and expression of the partner gene to similar extents. For 170 genes, antisense transcription was significantly correlated to both sense gene expression and splicing. A few genes had antisense-correlated changes only in splicing (21) or expression (6). As observed in previous studies (Chen et al. 2005; Riken Genome Exploration Research et al. 2005), the expression of most SAS gene pairs (96.6%) was positively correlated, indicating concordant expression.

3.2.5 Regions of SAS overlap are enriched in exons with antisense-correlated splicing events

RNA-masking of splice sites via dsRNA formation underlies antisense-mediated splicing regulation of genes such as *TR*α (Hastings et al. 1997), indicating the importance of sequence overlap. To determine the relative importance of SAS sequence overlap in our data, we ascertained whether exons that overlapped an antisense gene ("overlapping exons") were more likely to exhibit antisense-correlated splicing events than exons outside of the annotated overlap ("non-overlapping exons") (**Fig. 3.1A**).

Of the 191 known SAS genes with antisense-correlated splicing events, 27 had at least two overlapping and two non-overlapping expressed exons. For each of these genes, we compared the proportion of antisense-correlated non-overlapping exons to that of antisense-correlated overlapping exons. If sequence overlap is not an important factor, the proportions of these two groups should be equal. Instead, for 21 genes (77.8%), a greater proportion of exons with antisense-correlated splicing were overlapping rather than non-overlapping (**Fig. 3.5B**). Physical overlap therefore seems to be a critical aspect

of the observed antisense-correlated splicing events, perhaps indicating that sequence overlap is a key feature of the mechanism of splicing control acting at these loci.

3.2.6 Regions of SAS overlap are enriched in nucleosomes, PolII occupancy and alternatively spliced exons

Recent analyses (Nahkuri et al. 2009; Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009) of publicly available ChIP-seq data from human T-cells (Schones et al. 2008), found that nucleosome occupancy is elevated in exons relative to introns, and indicate that this enrichment decreases the rate of RNA Polymerase II (PolII) elongation (Schwartz et al. 2009). Indeed, nucleosomes constitute chromatin "roadblocks" slow the PolII elongation rate (Kulaeva et al. 2009), and slower PolII elongation rates have in turn been shown to increase the rate of alternative splicing (de la Mata et al. 2003). Thus one hypothesis for the observed increase in the rate of antisense-correlated alternative splicing events in SAS overlaps may involve a decreased polymerase speed in those regions. In support of this hypothesis, we noted that areas of SAS overlap contained exons encoded by two genes, leading to a potential enrichment of exons in overlapping regions. We ascertained the frequency (per kb) of Ensembl-annotated exons in SAS genes (Methods), and found a 7.2-fold increased exon/kb frequency in overlapping (3.1 exons/kb) versus non-overlapping regions (0.43 exons/kb; Welch's t-Test, $P < 2.2\times10^{-16}$). This finding would correspond to a greater frequency of nucleosomes in overlapping regions if exons were indeed enriched in nucleosomes, as expected from previous studies (Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009).

3.2.6.1 Nucleosomes are enriched in areas of SAS overlap

To confirm that SAS genes harbored more nucleosomes in exons than introns, we re-analyzed the publicly available activated T-cell ChIP-seq and microarray data (Schones et al. 2008). We used the T-cell microarray data to identify 8,627 expressed genes in activated T-cells. Of these, 189 belonged to the set of 2,995 known SAS genes, and a smaller subset of 122 genes had antisense-correlated splicing events and corresponding nucleosome occupancy ChIP-seq data. On average, exons had a 1.2-fold enrichment of nucleosome peaks compared to the introns of the same genes (Student's t-Test $P = 5.7\times10^{-9}$; **Fig. 3.5A**). We conclude that both exons and nucleosomes are enriched in areas of SAS overlap (modeled in **Fig. 3.6**).

3.2.6.2 Increased PolII occupancy in regions of SAS overlap

Increased nucleosome occupancy of SAS overlaps should lead to attenuated PolII elongation speed in these regions. Given the documented effects of decreased PolII speed on alternative splicing (de la Mata et al. 2003), we also expected an increased local frequency of alternatively spliced exons. To test these predictions, we ascertained the level of both PolII occupancy and of alternative splicing in regions of SAS overlap relative to non-overlapping regions, as described next.

3.2.6.2.1 PolII occupancy

PolII occupancy levels were analyzed using publicly available ChIP-seq data from one LCL (GM12878), generated as part of the ENCODE project (Consortium 2004). We sought to determine whether PolII peaks were enriched in regions of sequence overlap relative to flanking non-overlapping regions in individual SAS genes. PolII occupancy was used as a surrogate measure of PolII speed, since areas with stalled or slowly-moving complexes were more likely to be observed as bound by PolII in a ChIP-seq experiment than areas with fast moving PolII complexes. Thus, PolII peaks are expected to represent regions of DNA through which PolII exhibits slow elongation speeds. To assess PolII occupancy, areas with significant enrichment of signal over background (peaks) were enumerated independently in overlapping and non-overlapping regions of known SAS genes. A total of 549 known SAS genes were expressed in GM12878, and harboured at least one significant PolII peak. Of these, we analyzed 248 genes with at least one non promoter-associated PolII peak, indicating the presence of the elongating form of PolII in the gene body. These genes harbored 488 PolII peaks in distinct regions: 85 peaks (17.4%) in known promoters, 212 peaks (43.4%) in non-overlapping regions, and 191 peaks (39.1%) in overlapping regions. Regions of overlap spanned an average of 11.1% of the total gene lengths. By calculating the log ratio of overlapping versus non-overlapping PolII occupancy levels (peaks/kb), a 5.5 fold enrichment was observed in areas of overlap for 85.9% of the 248 known SAS genes (**Fig. 3.4C**, Mann-Whitney Test, $P = 2.4\times10^{-19}$). We excluded the possibility that enrichment of PolII peaks in areas of SAS overlap was due to transcriptional termination, which can also decrease PolII speed (Nag et al. 2006) (data not shown). This enrichment corresponds with the anticipated effect of increased nucleosome concentrations on PolII speed in areas of SAS overlaps, and is likely to cause local changes in splicing outcomes.

3.2.6.2.2 Higher rates of alternative splicing in areas of SAS overlap

To investigate changes in alternative splicing in overlapping versus non-overlapping regions, we identified constitutive and alternative exons for 8,530 Ensembl genes with multiple transcripts. The 149,032 exons encoded by these genes were categorized as "constitutive" if present in all annotated gene isoforms (45.5% of exons), and "alternative" if found in only a subset of isoforms (55.5% of exons). Next, all exons encoded in the 2,668 known SAS genes were subdivided into overlapping and non-overlapping categories, as previously described. Of these, we analyzed 163 genes that had both alternative and constitutive exons, and at least two overlapping and two non-overlapping exons. A total of 57.1% of non-overlapping exons were alternatively spliced, similar to the proportion of alternative exons in all 8,530 genes with multiple isoforms (**Table 3.1**, Student's t-Test, P = 0.6). When considering overlapping exons however, 67.8% of exons were alternatively spliced, a significant increase from the overall proportion (**Table 3.1**, Student's t-Test, $P = 4.5\times10^{-4}$). Elevated levels of alternative splicing were expected from the inferred local decrease in PolII transcriptional speed, and suggest that antisense transcription ultimately increases the variety of alternative transcripts expressed from SAS loci (modeled in **Fig. 3.6**).

3.2.7 Antisense transcription coincides with alternative splicing throughout metazoan evolution

Since antisense transcription has been observed in numerous organisms (Dahary et al. 2005); (Zhang et al. 2006), we hypothesized that the relationship between splicing and antisense transcription is evolutionarily conserved. To address this possibility, we measured the concordance between alternative splicing (inferred from the number of annotated sense gene isoforms) and annotated antisense genes in twelve species: human, mouse, rat, chimp, rhesus monkey, fly, chicken, frog, sea squirt, puffer fish, worm, and zebrafish. We first divided genes into those with multiple annotated isoforms and those with a single known transcript (**Fig. 3.7A**). In each species, we then compared the proportion of known SAS genes in each category, and found that a significantly higher proportion of multiple-transcript genes had known antisense gene partners in eleven species (**Fig. 3.7B**, corresponding P values in **Table 3.2A**).

We next measured novel antisense transcription by utilizing species-specific ESTs mapping antisense to known genes (Methods). Antisense ESTs were not only found in a significantly larger proportion of genes with multiple rather than single isoforms (**Fig. 3.7C**, **Table 3.2B**), but they were also more highly expressed, indicating that antisense transcription is abundant at these loci. Together, these findings indicate that antisense transcription is a general feature of genes with multiple transcripts throughout evolution.

3.3 Discussion

Numerous groups have reported on the abundance of antisense transcription in mammalian transcriptomes (Engstrom et al. 2006); (Riken Genome Exploration Research et al. 2005); (Chen et al. 2004; Kapranov et al. 2005) and on the frequent co-expression of SAS gene partners (Reis et al. 2004; Chen et al. 2005; Kiyosawa et al. 2005), but the functional implications of this transcription remain to be elucidated. For the first time we provide evidence linking antisense transcription to alternative splicing across the majority of expressed human SAS loci. First, antisense transcription distinguishes long genes with numerous exons and transcript isoforms from shorter genes with simpler splicing outcomes. Second, both known and novel instances of antisense transcription are strongly correlated to sense gene splicing, affecting 20-24% of exons at 74-79% of expressed known and novel SAS loci. Together, these findings provide a basis for interpreting potential functional outcomes of co-expressed SAS genes.

Altering the complement of proteins associated with the PolII C-terminal domain can affect splicing either by altering the elongation speed of the polymerase or by making specific splicing factors available co-transcriptionally (Listerman et al. 2006), thus affecting the alternative expression of many genes. In contrast to such trans-acting effects of classical splicing regulatory mechanisms (Wang and Burge 2008), a distinguishing aspect of antisense-mediated splicing regulation is its effect on individual *cis*-encoded genes, yet in a manner unique from that of other *cis*-acting elements, such as splicing enhancers (Castle et al. 2008). The importance of SAS sequence overlap underscores these key differences, since overlapping regions are enriched in antisense-correlated alternative splicing events. Our findings indicate that overlapping regions are characterized by a greater frequency of exons and elevated nucleosome occupancy compared to adjacent non-overlapping regions. The elevated nucleosomal barrier in these

regions is correlated to decreased PolII speed, which is further linked to an observable increase in alternative exon usage in areas of SAS overlap (**Fig. 3.5**). A similar increase in nucleosome occupancy has previously been linked to actively used polyadenylation signals (PAS) in T-cells (Spies et al. 2009), as well as decreased polymerase speed in intragenic regions (de la Mata et al. 2003; de la Mata and Kornblihtt 2006). In conjunction, these observations underscore the role that sequence-based determinants of nucleosome positioning, (such as nucleosome binding affinity of exonic and PAS-associated sequences) play in alternative polyadenylation and splicing.

In contrast to known SAS genes, the antisense transcripts at novel SAS loci do not correspond to known genes with identifiable exons, yet antisense-correlated splicing still occurs with a similar prevalence. At such loci, elongating polymerases transcribing the sense gene could be slowed either by increased nucleosome occupancy levels as seen in known SAS pairs, or alternatively, by the non-mutually exclusive mechanism of transcriptional interference, which has previously been shown to slow PolII speed (Galburt et al. 2007); (Shearwin et al. 2005). A significant amount of novel SAS transcription was detected in multiple species. Along with our observations of human population-specific differences in novel SAS events correlated to alternative splicing events, these results suggest that new SAS loci continue to evolve and to influence splicing outcomes.

We found a strong concordance between known antisense transcription and genes with multiple isoforms in amphibians, fishes, insects, birds, nematodes, and mammals. If antisense-mediated regulation of alternative splicing were functionally important in these species, we would expect SAS gene overlaps to be conserved throughout evolution. In support of this hypothesis, positive selection for the maintenance of sequence overlaps in known SAS genes has already been documented between the human, mouse, and Fugu genomes (Dahary et al. 2005). In conjunction with detectable alterations in chromatin-state and PolII processivity at human known SAS loci, these observations advocate for antisense transcription as a conserved mechanism of splicing regulation.

3.4 Methods

3.4.1 Ensembl genes

Ensembl (Hubbard et al. 2002) gene annotations (including gene, transcript, and exon coordinates; release 49) were downloaded via the Ensembl Perl API. Genes whose genomic coordinates overlapped by at least one base, and which were encoded on opposite strands were categorized as known SAS genes. Exons were classified as alternative (A) or constitutive (C) if they were found in a subset or in all of the annotated isoforms of a gene, respectively. Only known SAS genes with both A and C exons, and at least two expressed exons in the overlapping and two expressed exons in the non-overlapping SAS region were considered. Affymetrix probesets were mapped to the sense and antisense strands of genes using custom Perl scripts.

3.4.2 Public datasets

3.4.2.1 Lymphoblastoid cell lines

Publicly available CEU and YRI Affymetrix Human Exon 1.0 ST Array data (http://media.affymetrix.com:80/support/technical/technotes/exon_array_design_technote.pdf) were downloaded from the Gene Expression Omnibus (Barrett et al. 2009) (GEO, GSE7792). A total of 18,041 genes had probesets mapping to both the positive and negative strands of the genome, and 10,636 genes had probesets mapping only to the sense strand. An additional 366 genes had probesets mapping only to the antisense strand, and likely reflect changes in gene annotations since probeset design.

3.4.2.2 Multiple species data

Current gene annotations and EST data were downloaded from the UCSC genome browser (Rosenbloom et al. 2007) for human (*Homo sapiens*), puffer fish (*Takifugu rubripes*), mouse (*Mus musculus*), chimp (*Pan troglodytes*), rhesus (*Macaca mulatta*), rat (*Rattus norvegicus*), sea squirt (*Ciona intestinalis*), fly (*Drosophila melanogaster*), frog (*Xenopus tropicalis*), chicken (*Gallus gallus*), nematode (*Caenorhabditis elegans*), and zebrafish (*Danio rerio*).

3.4.2.3 PolII data

ChIP-seq data were downloaded from the UCSC genome browser (Primary Table: *wgEncodeYaleChIPseqRel2SignalGm12878Pol2*). For the PolII analysis, known SAS genes were required to have at least one PolII peak in the gene body (i.e. not including the first exon of the gene). Genes completely overlapped by another gene were excluded from the analysis.

3.4.2.4 Nucleosome data

Activated T-cell ChIP-seq and microarray expression data were downloaded from GEO (Barrett et al. 2009) (GSE10437). T-cell microarray data (Schones et al. 2008) were processed using the Affymetrix Expression Console Software (http://www.affymetrix.com/), and MAS5 (Hubbell et al. 2002) p-values were used to identify expressed genes. ChIP-seq data were processed as previously described (Schwartz et al. 2009). Mean nucleosome occupancy was calculated separately for intronic and exonic regions in SAS gene partner regions. Areas of exon-intron sequence overlap were considered exonic sequence.

3.4.2.5 Cancer Gene Census data

CGC genes were obtained from the Wellcome Trust Sanger Institute Cancer Genome Project web site, http://www.sanger.ac.uk/genetics/CGP.

3.4.3 Microarray processing

3.4.3.1 Pre-processing

Array data were background corrected and normalized according to standard protocols (Expression Console, www.affymetrix.com/support/technical/software_downloads.affx). The $\log_2$ of the resulting expression values was used in further analyses.

3.4.3.2 Gene expression filtering

Probesets were filtered for expression above background (Griffith et al. 2008) in at least 20% of samples and gene-level expression values were calculated for genes that had a minimum of 20% of probesets expressed in at least 20% of samples, and a minimum of two expressed probesets. In the case of novel SAS genes, an “antisense construct” was generated to represent the unknown antisense transcript. The boundaries of the antisense

construct were set to the genomic boundaries of the sense gene, but only probesets mapping to the opposite strands were considered (Fig. 1B). Probesets mapping in this region were used to calculate the antisense construct expression in an analogous way to annotated genes.

3.4.3 Splice index calculations

Gene expression was calculated as the mean of all probesets mapping to the sense strand of that gene or antisense construct. Probesets that mapped to introns as well as exons were included, since they may represent alternatively spliced exons, intron retention events, or other un-annotated splicing variations, such as alternative 5′ or 3′ splice-site usage. Each probeset is therefore referred to as an exon. The splice index was the expression of the exon normalized to the expression of the whole gene:

$$\textit{Splice index (exon)} = \textit{expression (exon)} / \textit{expression (gene)}$$

The Spearman's rank correlation coefficient of each sense exon splicing index and the antisense gene (or construct) expression was calculated for all samples and all SAS genes, using the cor.test function in R (R_Development_Core_Team 2008). Associated correlation p-values (Best and Roberts 1975) were multiple-test corrected using the Bonferroni method (Wright 1992). In known SAS gene pairs, each gene partner was in turn analyzed as the sense gene and as the antisense gene. Correlations (and associated p-values) between gene expression values were calculated using the same methods.

Relative to probesets that were not antisense-correlated, correlated probesets did not have biases in any of the following features: number of independent probes, cross-hybridization type, or probe count (Chi-square test, respective P values = 0.98, 0.80, 1.00).

3.4.4 Functional annotation

We used the David Bioinformatics Resources (Dennis et al. 2003); (Huang et al. 2009) to functionally annotate the 4,792 known SAS, 7,648 novel SAS and 7,137 non SAS genes whose IDs could be converted to DAVID IDs, specifically focusing on Gene Ontology (Michael et al. 2000) terms (GO), and Uniprot Keywords (http://www.uniprot.org/manual/keywords).

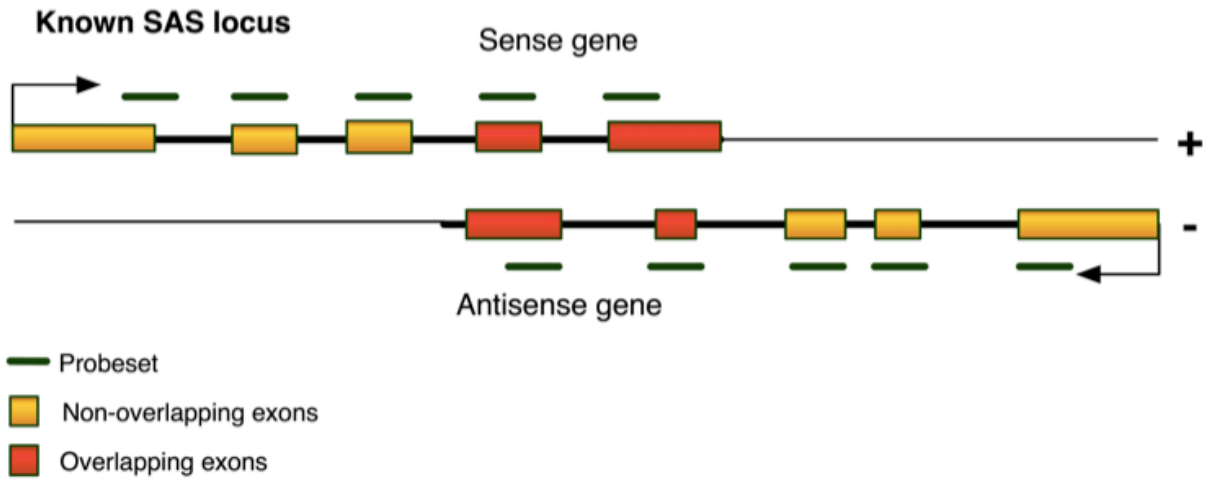
3.4.5 Exon frequency calculations

The frequency of exons per kilobase (exons/kb) was calculated for 1,765 known SAS gene pairs. For each gene pair, the number of exons/kb in the overlapping region (including exons from both strands) was compared to the number of exons/kb in non-overlapping regions of both genes. For this analysis, overlapping alternative exons (ie. sharing the same genomic location, but differing in 5′ or 3′ ends) were only counted once.

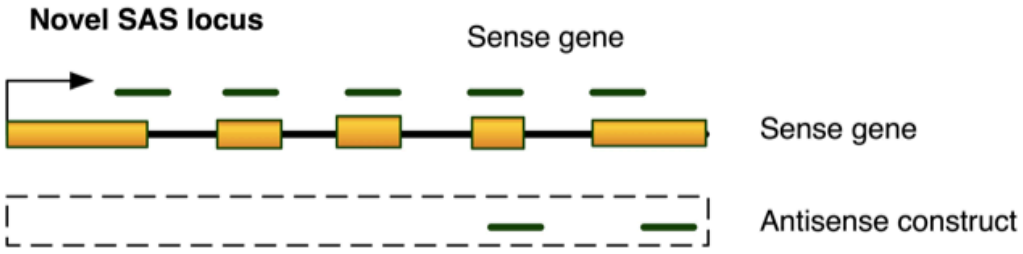
Figure 3.1 Known and novel SAS genes are structurally distinct from genes without antisense transcription.

(A)**.** Schematic diagram of a hypothetical known SAS gene pair shows the structural arrangement of overlapping exons (red rectangles) and non-overlapping exons (orange rectangles). In our analysis, each partner gene is in turn treated as the antisense gene. Probesets (horizontal green dashes) map to the sense strand of the gene in either exons or introns. Black arrows denote transcriptional direction. (B) A novel SAS sense gene. Since the structure of the antisense transcript is unknown, an antisense construct (dashed-line box) spans the genomic coordinates of the sense gene and approximates antisense expression. All antisense probesets encompassed by that region are used to infer the expression of the antisense construct. If the actual antisense transcript extends beyond the sense gene boundaries, the antisense construct expression under-represents the actual level of antisense transcription at that locus. (C). On average, known and novel SAS genes have significantly longer gene (left vertical axis) and transcript lengths than genes with no antisense transcription, as well as significantly more introns and isoforms (right vertical axis). The total number of protein coding known SAS genes, novel SAS, and non SAS genes in the human genome is shown on the x-axis. (D) P-values for all pairwise comparisons between known SAS, novel SAS, and non SAS genes (Welch two-sample t-test).

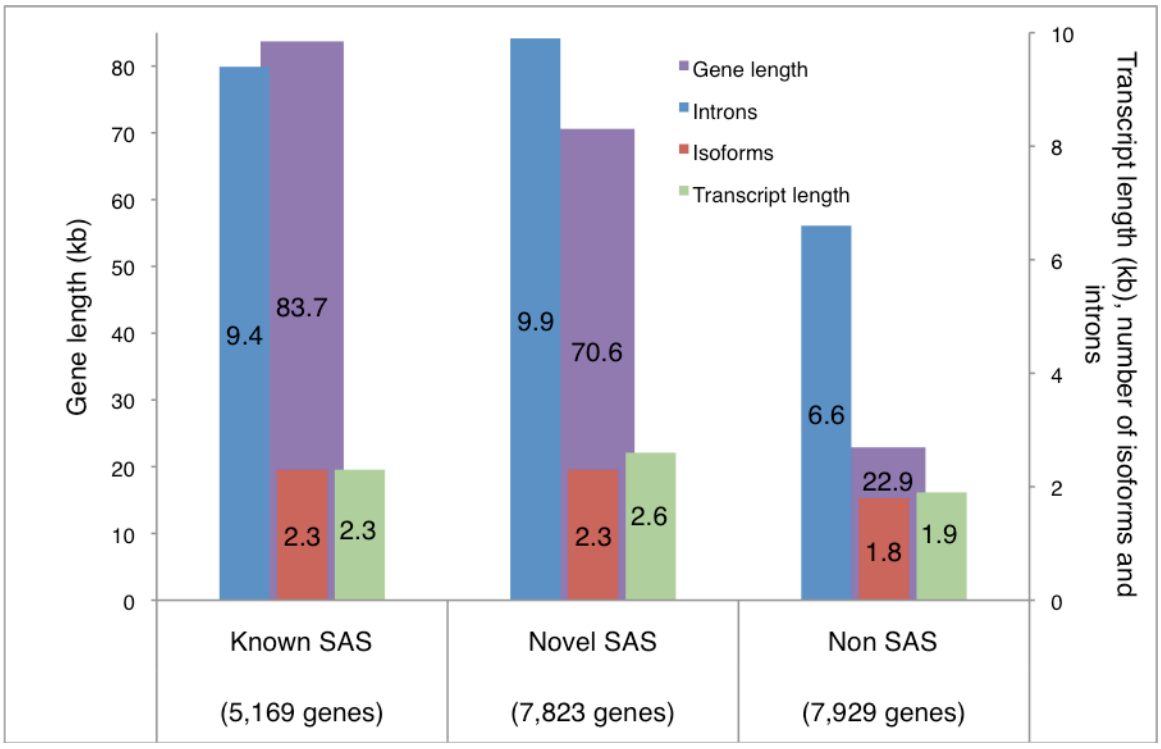
A


Known SAS locus
Sense gene
+
-
Antisense gene
Probeset
Non-overlapping exons
Overlapping exons

B


Novel SAS locus
Sense gene
Sense gene
Antisense construct

C



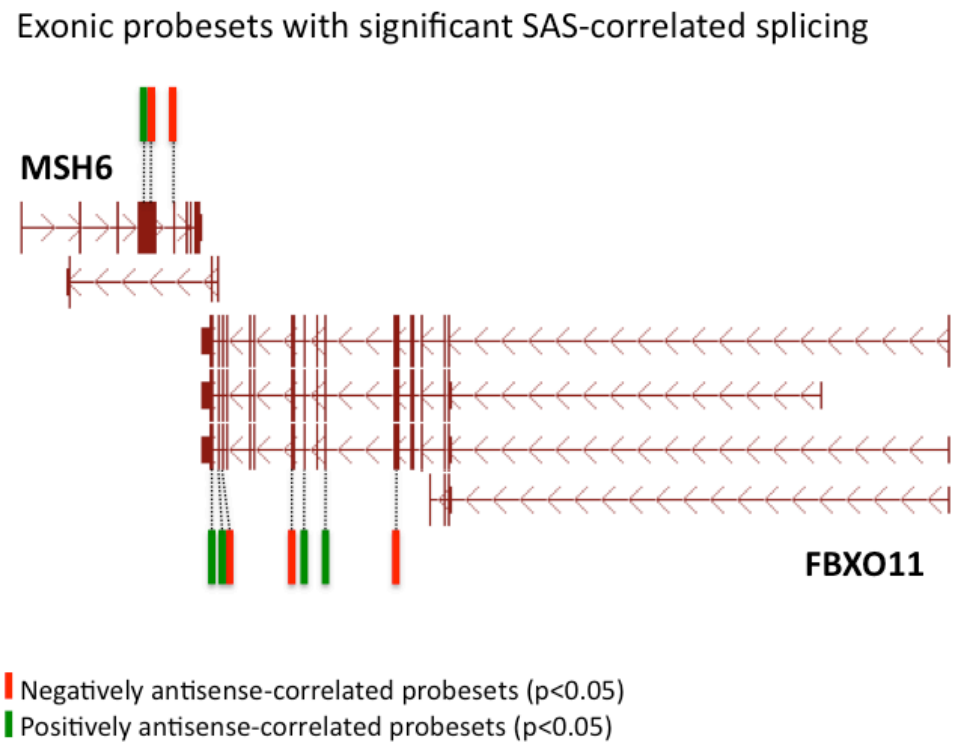
D

	Known SAS vs Novel SAS	**Known SAS vs Non SAS**	**Novel SAS vs Non SAS**
Gene length (kb)	2.4×10^{-7}	1.6×10^{-148}	1.0×10^{-285}
Transcript length (kb)	8.0×10^{-18}	5.1×10^{-40}	3.0×10^{-147}
Transcripts per gene	6.1×10^{-1}	3.2×10^{-84}	4.2×10^{-111}
Introns per gene	5.8×10^{-3}	5.8×10^{-78}	9.7×10^{-147}

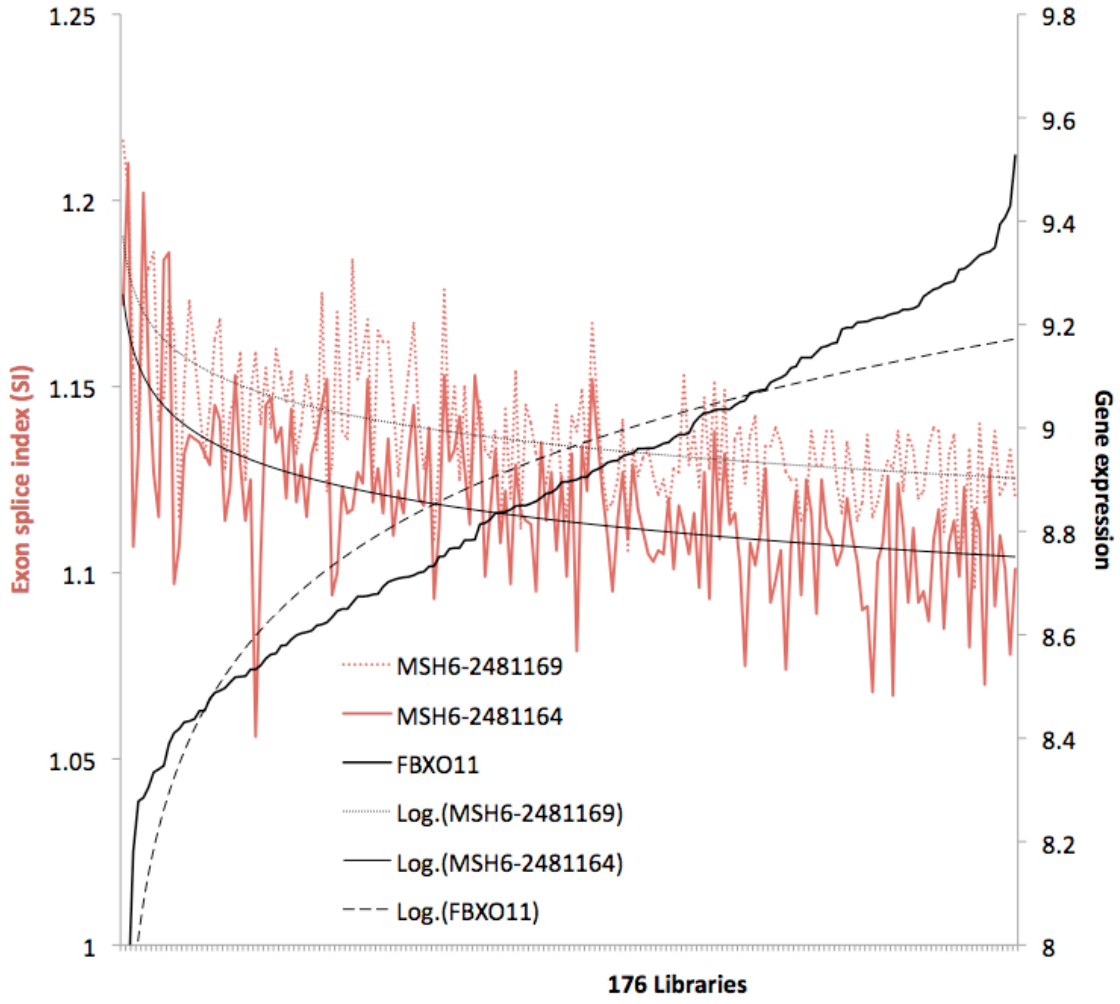
Figure 3.2 Antisense-correlated splicing events at the MSH6 and FBXO11 locus.

(A) EnsEMBL isoforms of MSH6 and FBXO11 are shown as exons (burgundy rectangles) separated by introns (horizontal lines), along with arrows denoting transcriptional direction. The gene MSH6 largely overlaps the FBXO11 gene, and has one exonic probeset whose splice index is positively antisense-correlated (green rectangle), and two exonic probesetss whose splice index is negatively antisense-correlated (red rectangles). A black dotted line links each probeset with the exon it represents. The SAS partner gene (FBXO11) has four exons with positive antisense-correlated splicing, and three exons with negative antisense-correlated splicing. (B) The splice index values of the negatively antisense-correlated MSH6 exons (red solid and red dashed lines, left-hand y-axis) are shown along with the expression of the antisense gene (FBXO11, solid black line, right-hand y-axis). Values are reported for all 176 CEPH and YRI libraries (x-axis). Trendlines (log-transformed) were fitted to the gene and exon values. (C) The splice index values of the positively correlated MSH6 exon (green solid line) is shown along with the FBXO11 antisense gene expression. As in (B), trendlines are shown for each dataset. In both (B) and (C), the libraries are sorted by ascending FBXO11 expression values.

A



B


1.25
1.2
1.15
1.1
1.05
1
Exon splice index (SI)
9.8
9.6
9.4
9.2
9
8.8
8.6
8.4
8.2
8
Gene expression
MSH6-2481169
MSH6-2481164
FBXO11
Log.(MSH6-2481169)
Log.(MSH6-2481164)
Log.(FBXO11)
176 Libraries

C

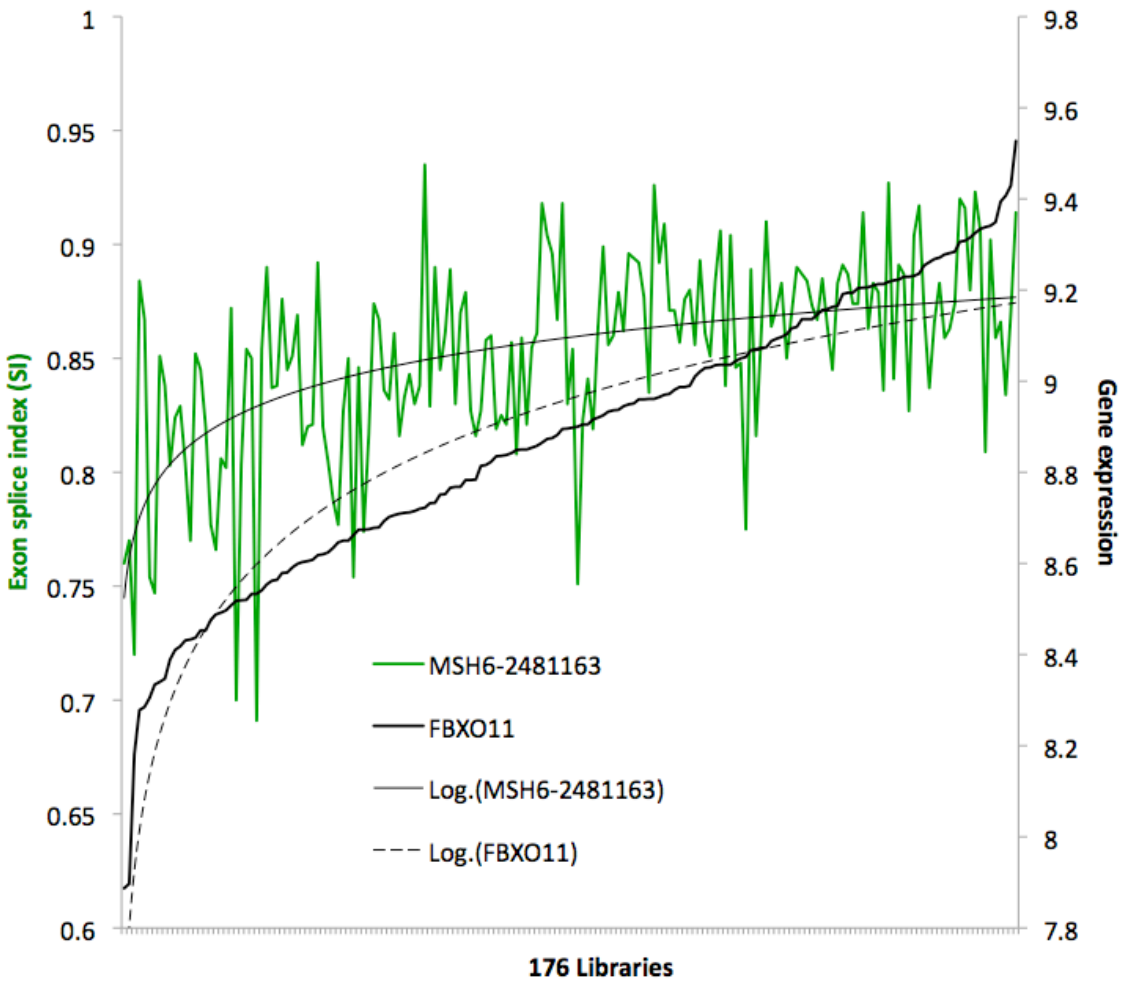

1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Exon splice index (SI)
9.8
9.6
9.4
9.2
9
8.8
8.6
8.4
8.2
8
7.8
Gene expression
MSH6-2481163
FBXO11
Log.(MSH6-2481163)
Log.(FBXO11)
176 Libraries

Figure 3.3 The majority of antisense expression is significantly correlated to sense gene exon splicing events.

Dashed arrows mark significant correlations between the splicing index of individual sense-gene exons and the expression of the antisense gene or construct (encased in dashed-line box). Significantly antisense-correlated exons are denoted by small orange rectangles; uncorrelated exons are small blue rectangles. The percent of expressed genes with antisense-correlated splicing events are shown above the diagram, while the percent of expressed exons whose splicing is correlated to antisense expression is shown below the diagram.

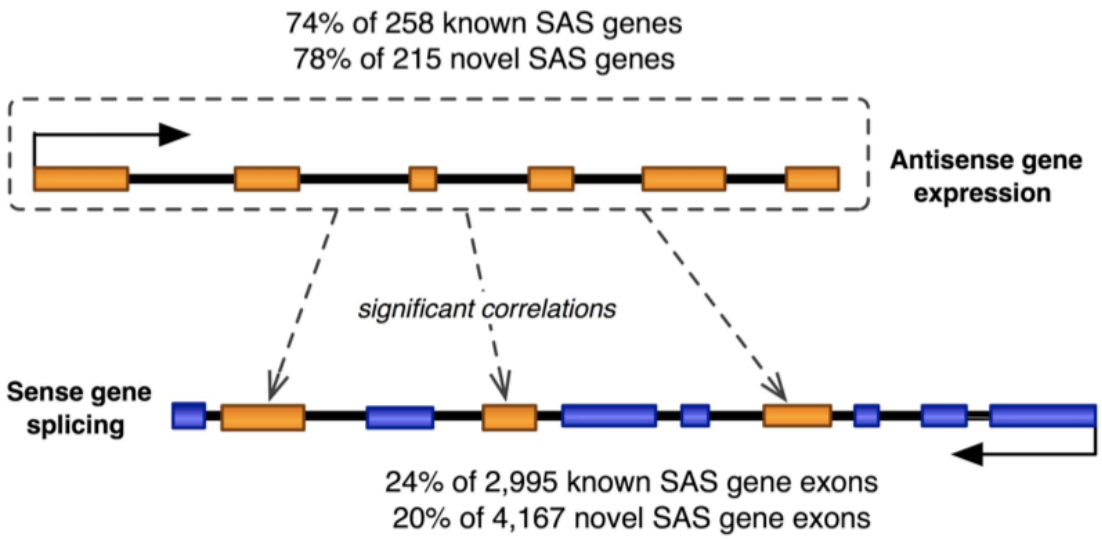
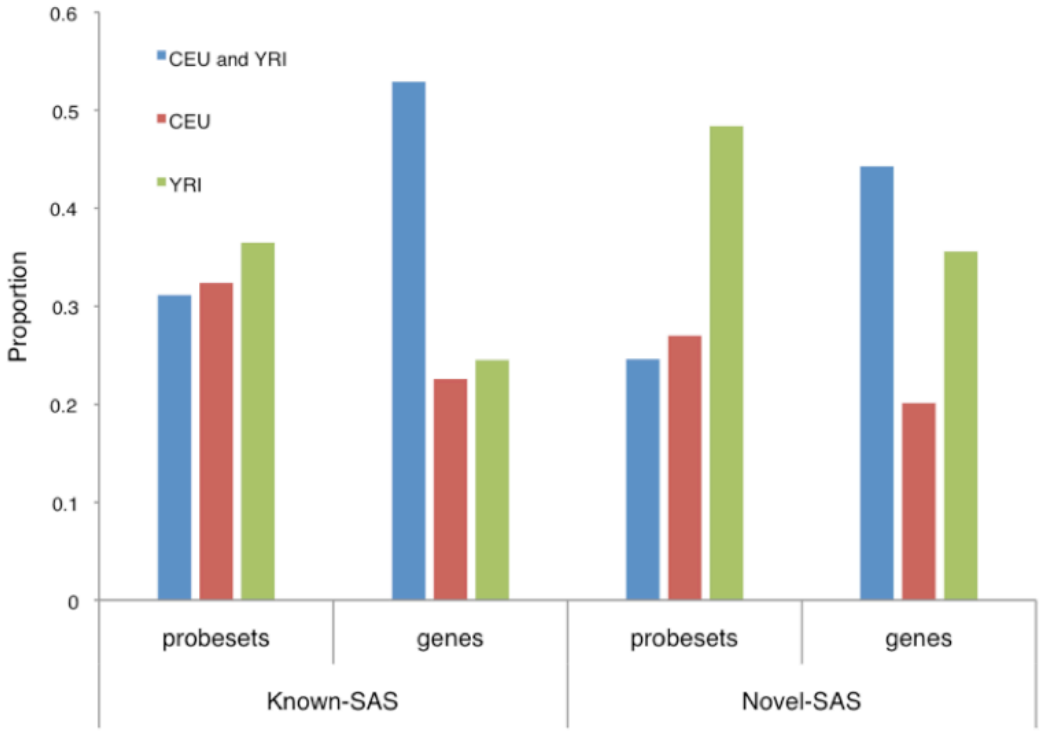


Figure 3.4 YRI individuals have a greater proportion of novel SAS genes with antisense-correlated splicing events.

The fraction (A) and number (B) of probesets (i.e. exons) with antisense-correlated splicing indexes, and the fraction of genes that these probesets map to, is shown for the individual CEU and YRI datasets and for both populations.

A



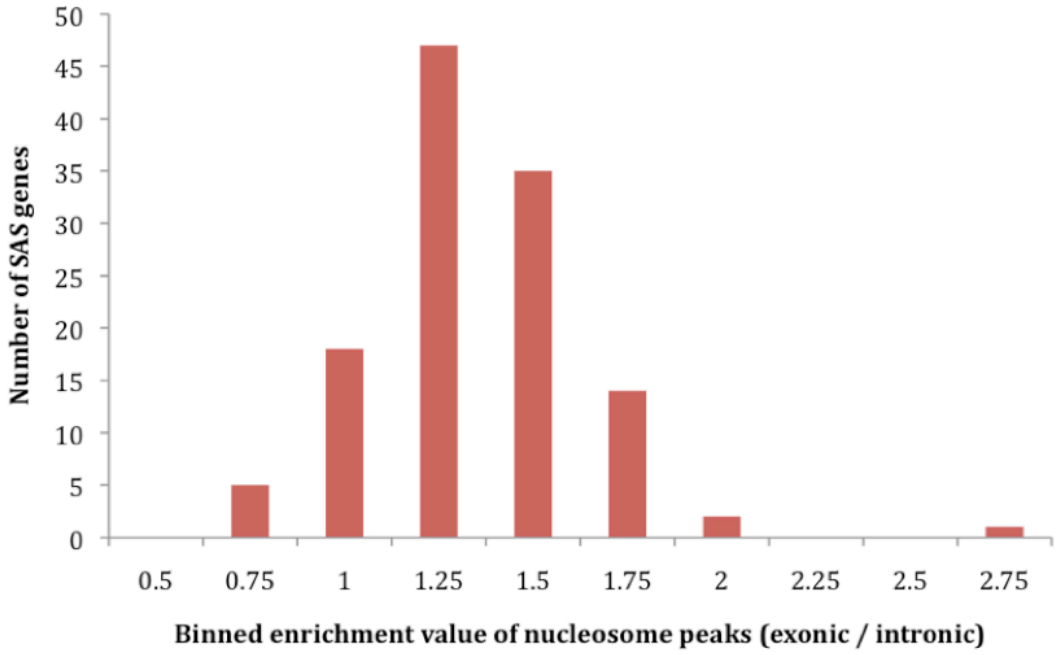
B

	Known SAS		Novel SAS	
	probesets	**genes**	**probesets**	**genes**
CEU and YRI	151	82	162	66
CEU	157	35	178	30
YRI	177	38	319	53
Total	485	155	659	149

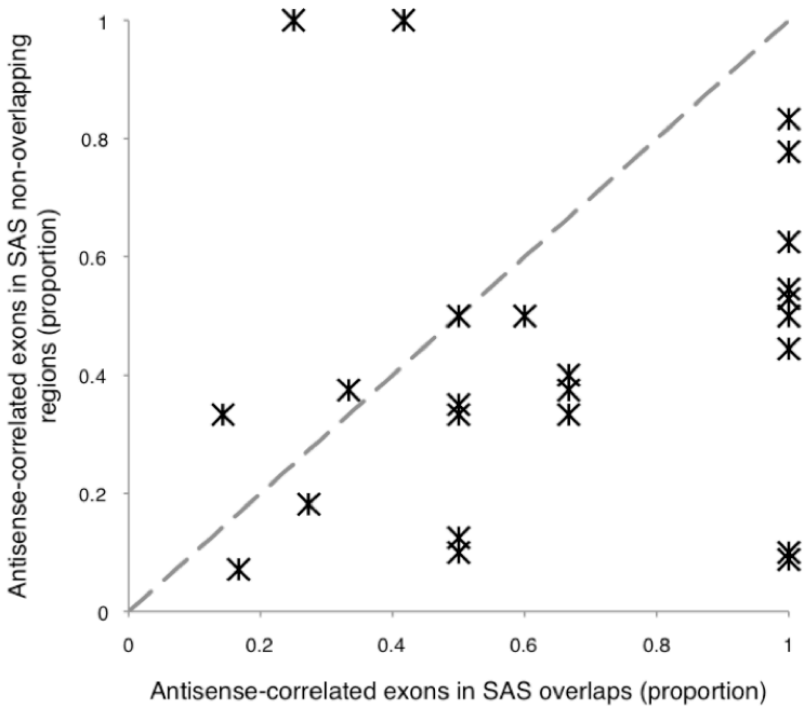
Figure 3.5 Nucleosomes, antisense-correlated splicing events, and PolII occupancy levels are enriched in SAS overlaps.

(A) Nucleosome peak enrichment was calculated using a ratio of exonic vs intronic peaks. The majority of genes have a significant enrichment of exonic nucleosomes in introns (ratio > 1). (B) The x-axis coordinate of each gene shows the fraction of all expressed overlapping exons with antisense-correlated splicing. The y-axis shows the fraction of all expressed non-overlapping exons with antisense-correlated splicing. The grey (dotted) line represents equal proportions of antisense-correlated overlapping and non-overlapping exons (C). The $\log_{10}$ ratios of overlapping versus non-overlapping PolII peaks/kb for 248 known SAS genes reveals PolII enrichment in SAS overlaps (log ratios > 0) at the majority of genes.

A



B


Antisense-correlated exons in SAS non-overlapping regions (proportion)
Antisense-correlated exons in SAS overlaps (proportion)
0
0.2
0.4
0.6
0.8
1

C

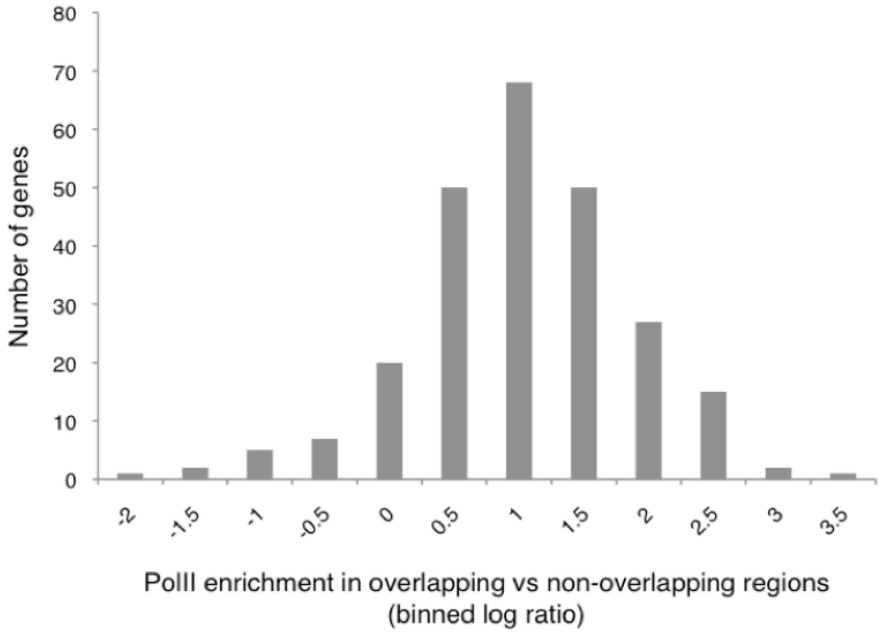

Number of genes
PolII enrichment in overlapping vs non-overlapping regions
(binned log ratio)
0
10
20
30
40
50
60
70
80
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5

Figure 3.6 Model of distinct features enriched in SAS overlapping regions.

Features over-represented in the SAS overlap (large rectangle) include exon frequency (blue or green rectangles connected by a thick black line), the proportion of alternative (green) versus constitutive (blue) exons, PolII peak frequency (orange rectangles), and the proportion of exons with antisense-correlated splicing patterns (*). Nucleosomes (gray ovals) are localized to exons, and are therefore enriched in the area of SAS overlap. Black arrows denote transcriptional direction.

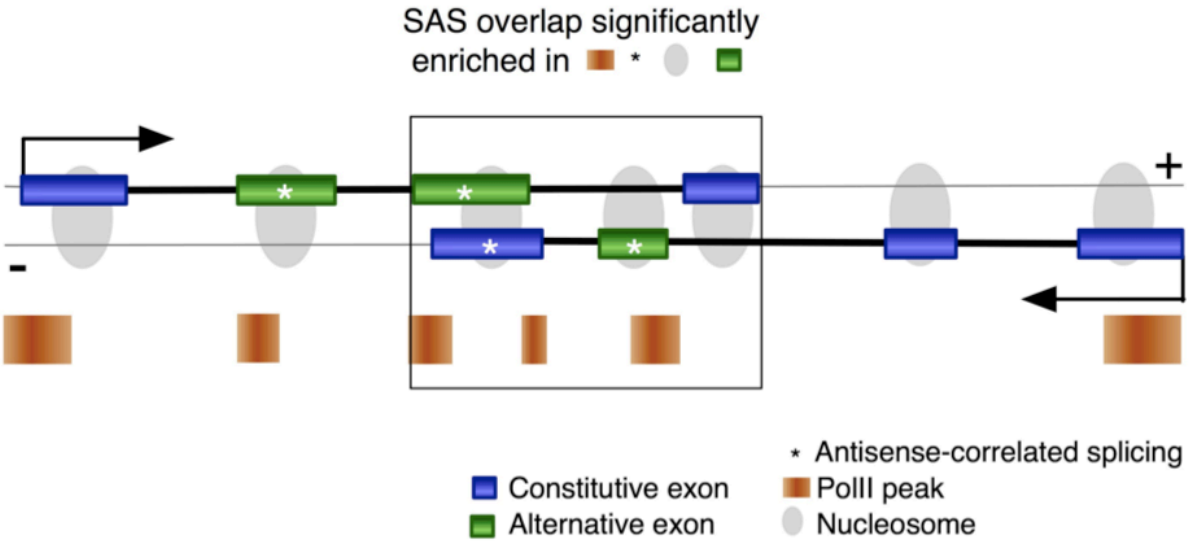
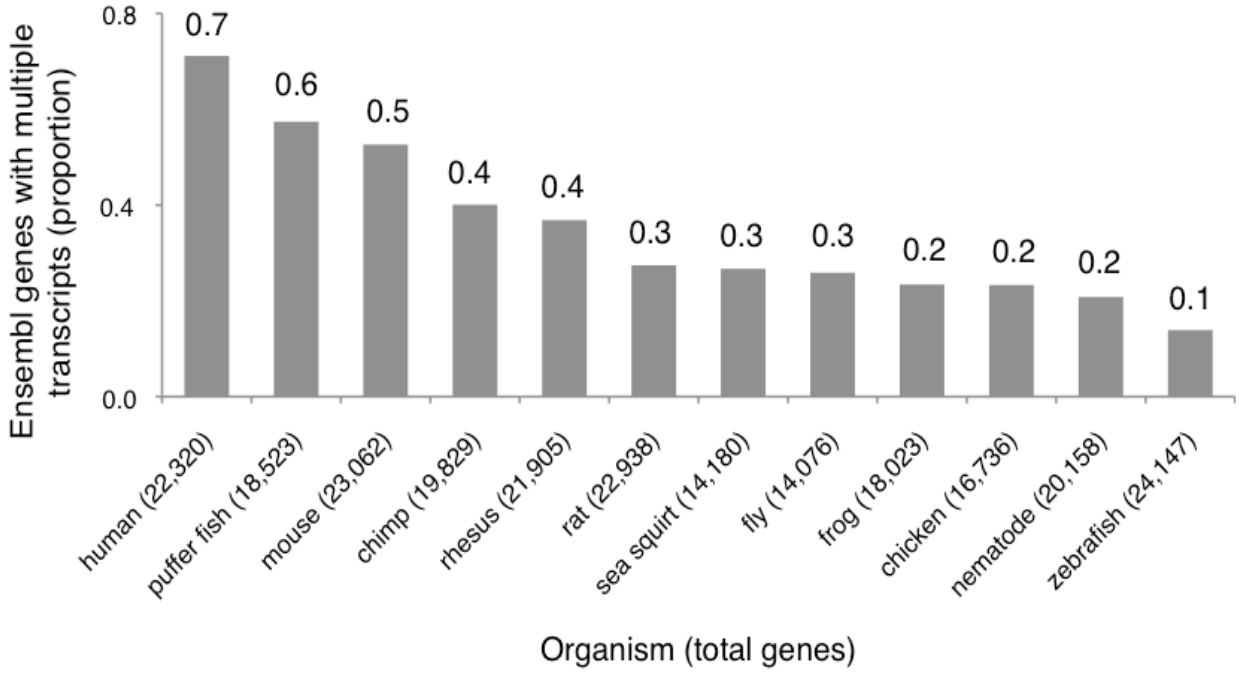


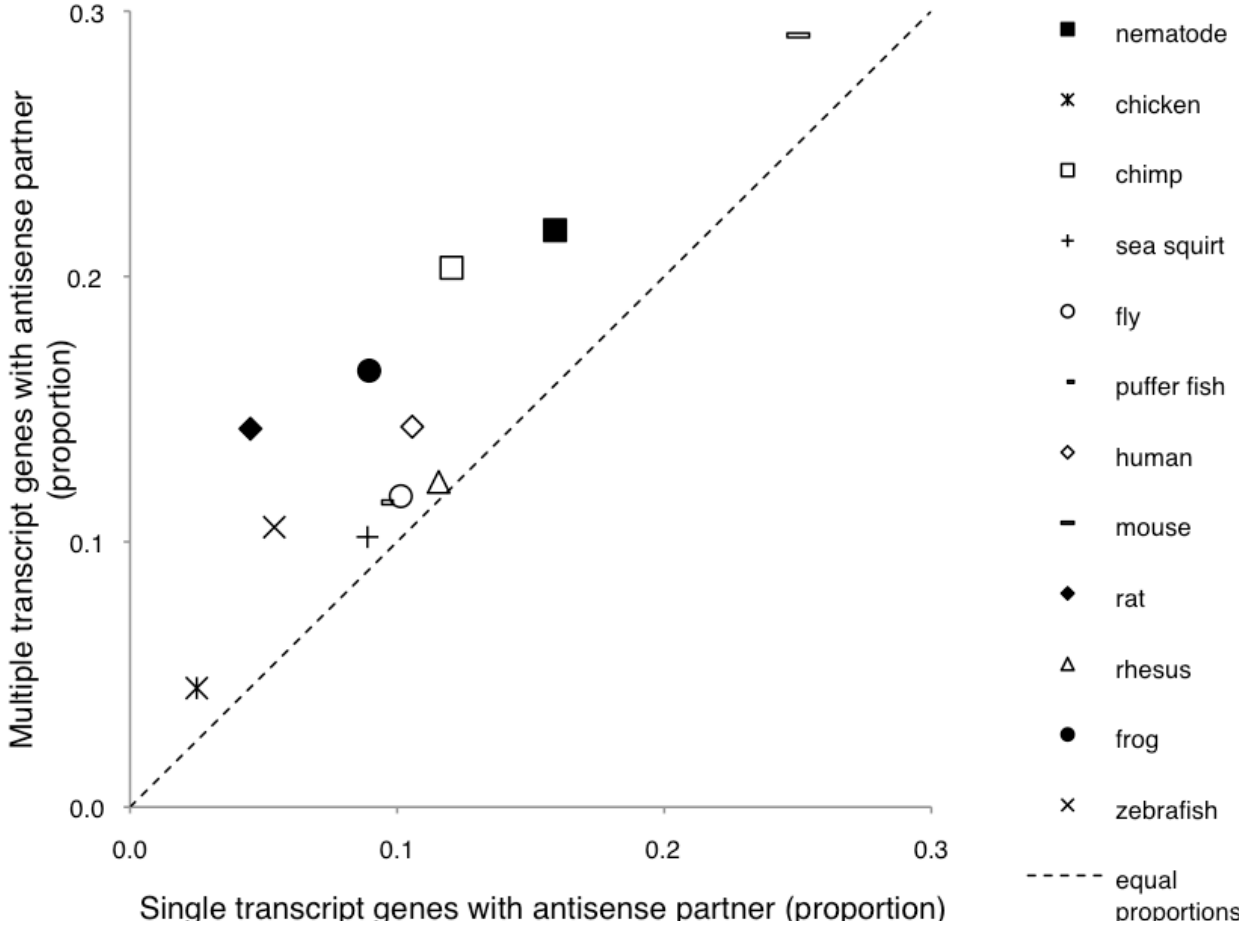
Figure 3.7 High concordance between alternative splicing and antisense transcription in multiple species.

(A) The proportion of all genes with multiple isoforms in twelve species. (B) Genes with multiple isoforms are enriched in known SAS pairs and (C) in EST evidence for novel antisense transcription (novel SAS genes). The dotted lines represent equal proportions of SAS genes or antisense ESTs among genes with multiple or single transcripts.

A



B


0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
Multiple transcript genes with antisense partner (proportion)
Single transcript genes with antisense partner (proportion)
nematode
chicken
chimp
sea squirt
fly
puffer fish
human
mouse
rat
rhesus
frog
zebrafish
equal proportions

C

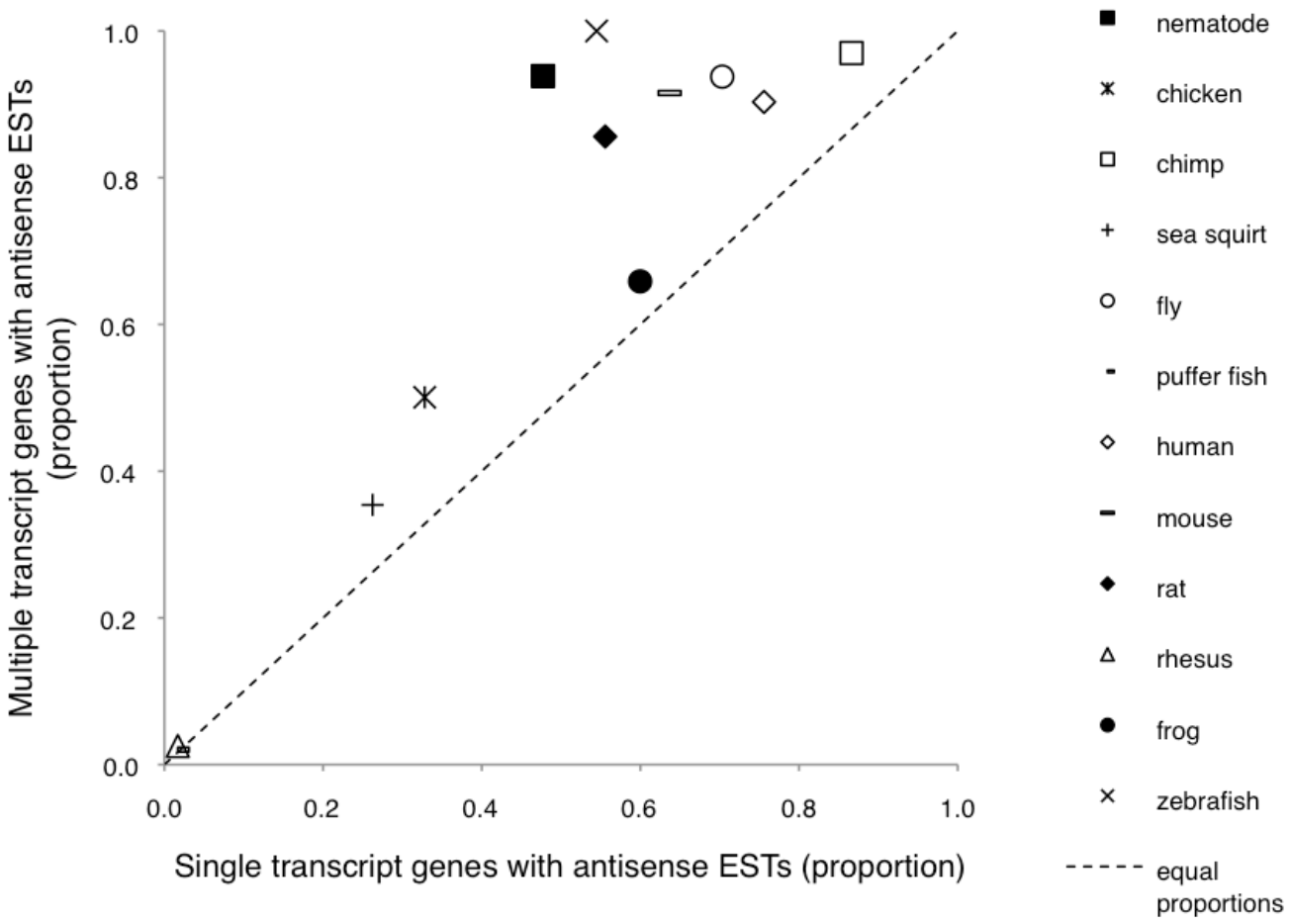

Multiple transcript genes with antisense ESTs (proportion)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Single transcript genes with antisense ESTs (proportion)
nematode
chicken
chimp
sea squirt
fly
puffer fish
human
mouse
rat
rhesus
frog
zebrafish
equal proportions

Table 3.1 Alternative exons are enriched in SAS overlaps.

The percent of alternatively (A) and constitutively (C) spliced exons is shown for 8,530 Ensembl genes with multiple isoforms and encoding both A and C exons. The proportion of A and C exons in overlapping and non-overlapping regions is summarized for a subset of 163 known SAS genes. P values correspond to differences between the proportions of A and C exons in overlapping or non-overlapping regions versus the proportion of in all genes (Student's T-test).

	Genes	Alternative (%)	Constitutive (%)	P value
SAS gene overlapping regions	163	67.8	32.2	4.5×10^{-4}
SAS gene non-overlapping regions	163	57.1	42.9	0.61
All genes	8,530	55.5	44.5	NA

Table 3.2 Significant concordance of SAS genes with alternative splicing across species

For twelve metazoan organisms, a significant enrichment of known SAS genes, and novel antisense transcription (as measured by the presence of antisense ESTs) was observed in genes with multiple rather than single transcripts. (A) For each species, the proportion of genes with multiple or single isoforms that are known SAS genes is tabulated along with the enrichment of SAS genes in the multiple transcript genes, and the p-value of that enrichment (Student's t-Test). Similarly (B), the average expression of ESTs mapping antisense to genes with multiple or single transcripts is enumerated along with the enrichment and significance for each species.

A

	Known SAS (proportion)			
Species	**Genes with multiple isoforms**	**Genes with single isoforms**	**Enrichment**	**P value**
human	0.22	0.16	1.37	3.48×10^{-25}
puffer fish	0.04	0.02	1.80	9.64×10^{-14}
mouse	0.20	0.12	1.69	2.25×10^{-66}
chimp	0.10	0.09	1.14	2.61×10^{-03}
rhesus	0.12	0.10	1.16	3.35×10^{-04}
rat	0.11	0.09	1.22	8.97×10^{-06}
sea squirt	0.14	0.11	1.36	5.15×10^{-09}
fly	0.29	0.25	1.16	2.51×10^{-06}
frog	0.14	0.05	3.16	1.24×10^{-64}
chicken	0.12	0.12	1.06	2.41×10^{-01}
nematode	0.16	0.09	1.84	1.05×10^{-33}
zebrafish	0.11	0.05	1.95	2.12×10^{-20}

B

Species	Antisense ESTs (average) Genes with multiple isoforms	Genes with single isoforms	Enrichment	P value
nematode	22.71	3.81	5.97	3.3×10^{-90}
chicken	5.58	3.43	1.63	3.5×10^{-08}
chimp	119.07	69.97	1.70	2.2×10^{-80}
sea squirt	19.01	18.24	1.04	9.0×10^{-01}
fly	18.11	6.89	2.63	1.5×10^{-81}
puffer fish	0.06	0.04	1.49	3.8×10^{-10}
human	123.06	57.15	2.15	7.3×10^{-126}
mouse	41.74	18.85	2.21	1.1×10^{-95}
rat	18.53	10.23	1.81	8.3×10^{-38}
rhesus	0.17	0.14	1.23	1.0×10^{-01}
frog	18.72	18.71	1.00	9.9×10^{-01}
zebrafish	16.58	10.51	1.58	4.6×10^{-06}

4. Analysis of antisense-correlated splicing events in Glioblastoma Multiforme reveals subtypes of cancer

Author contributions

A.S.M. and M.A.M. conceived the analyses. A.S.M. designed and performed all computational analyses, created the figures and tables, and wrote the text.

4.1 Introduction

Alternative expression is a general feature of human gene expression, specifically acting to increase protein diversity from the majority of genes (Wang et al. 2008a; Licatalosi and Darnell 2010). In particular, tissue-specific and temporal patterns of alternative splicing underscore the importance of this phenomenon to mammalian development. Its mis-regulation can lead to disease (Hutton et al. 1998; Garcia-Blanco et al. 2004), and can play a role in cancer (reviewed in (Venables 2004)).

Antisense transcription is a prevalent feature in the human genome, and in a small number of cases has been shown to regulate the alternative expression of cis-encoded sense transcripts (**Chapter 1**). Antisense transcripts have been shown to control the expression of disease-causing genes (Tufarelli et al. 2003), to be differentially expressed between normal and cancerous states (Morrissy et al. 2009), and to correlate to splicing outcomes (**Chapter 3**; (Hastings et al. 2000; Morrissy et al. 2010a)).

Given the relevance of alternative expression to cancer biology, and the relationship between antisense transcription and alternative expression, I hypothesized that antisense transcription can provide insight into properties of cancer biology. I chose to explore this question through the analysis of exon-level expression data collected by The Cancer Genome Atlas (TCGA) Research Network as part of a comprehensive catalog of cancer genomic profiles (TCGA 2008). This effort has led to the detailed profiling of a large cohort of glioblastoma multiforme (GBM) patients. In adults, GBM is the most common form of malignant brain tumor, and is characterized by a dismal one-year median survival rate (Ohgaki and Kleihues 2007). To date, analysis of TCGA data including copy number variations, gene expression levels, and somatic mutations has pinpointed

specific signaling pathways that likely drive GBM tumorigenesis. Given its uniform clinical manifestation and extremely poor prognosis, current research efforts are aimed at finding molecular signatures for GBM subtypes that can be used as prognostic, or even therapeutic markers.

I investigated the relationship between alternative splicing and antisense transcription in a panel of TCGA GBM samples (TCGA 2008). My aims were to (1) discern the extent of cancer-specific antisense-correlated splicing events, and (2) to determine whether antisense-correlated splicing events could be used to identify clinically distinct subgroups of GBM patients.

4.2 Results

4.2.1 Prevalent antisense-correlated splicing events in cancers and normal tissues

To study antisense-correlated alternative splicing in normal and malignant tissues, I compiled a dataset of 230 publicly available Affymetrix Exon 1.0 ST arrays profiling expression of normal tissue samples (Barrett et al. 2009), and 266 arrays profiling glioblastoma (GBM) samples. An additional 518 ovarian cystadenocarcinoma (OVC) samples were also collected to allow later discrimination between cancer-specific and GBM-specific splicing events (**Table 4.1,** (TCGA 2008)). Arrays were pre-processed and filtered as described in Methods. As previously noted (Morrissy et al. 2010a), probesets are referred to as exons, since they generally represent exonic sequences; probesets mapping to introns may represent either novel exons, alternative acceptor and donor sites of annotated exons, or retained introns. The expression of filtered exons was used to calculate gene expression and splice indexes (SI) for all expressed genes and exons, in the GBM, OVC and Normal datasets independently. SI values represent the proportion of a gene's expression that is due to a particular exon, and changes in the SI value indicate changes in the relative inclusion or exclusion of an exon in the mRNA (ex. **Fig 4.1**). Sense genes with annotated antisense partners (known SAS) and sense genes with no annotated antisense partners but expression-based evidence for novel antisense transcription (novel SAS) were retained for further analysis (**Fig 4.2A**).

A total of 3,312 (69.0%), 2,179 (47.4%), and 3,099 (65.4%) expressed genes harbored antisense-correlated splicing events in the Normal, GBM, and OVC samples,

respectively (**Fig 4.2A)**. In each respective dataset, a total of 17,420 (16.3%), 9,410 (11.2%), and 14,610 (16.2%) exons were spliced in an antisense-correlated manner (**Fig 4.2A).**

The majority of genes (1,730, 79.4%) with antisense-correlated alternative splicing in GBM had at least one GBM-specific splicing event, and consequently at least one GBM-specific isoform (**Fig 4.2B**). A subset of 4,689 (49.6%) of the alternatively spliced exons was only observed in GBM samples (**Fig 4.2B**). Similarily, large proportions of isoforms were unique to each dataset; I thus hypothesized that their relative expression (ie. SI values) could be used to molecularly distinguish distinct normal tissues and GBM subtypes.

4.2.2 Unsupervised hierarchical clustering identifies known normal tissues

I performed unsupervised hierarchical clustering of the SI values of the 17,420 exons with significant antisense-correlated splicing events in the Normal dataset, and found distinct clusters demarcating individual tissues (**Fig. 4.3**). Similar tissues profiled in different labs, such as the brain tissues (fetal brain, adult brain, GBM controls, and cerebellum from the Affymetrix dataset; see **Table 4.1**) clustered together, indicating that lab and batch-specific biases were successfully removed during data normalization. As expected from this observation, the Affymetrix cerebellum sample clustered with the remaining brain samples rather than the other tissues it was profiled with. Similarly, all blood tissue samples clustered together despite being generated by different labs. I conclude that antisense-correlated exons have SI values that represent tissue-specific patterns of expression, and consequently provide an effective means of clustering biologically distinct samples.

4.2.3 Clinically relevant GBM subclasses identified using unsupervised methods

Having determined that SI values of exons with antisense-correlated splicing can be used to distinguish normal tissues, I next conducted unsupervised hierarchical clustering on GBM samples. The 1,000 most variant exons with cancer-specific expression were used for this analysis, along with 245 GBM samples for which clinical data was available. These 1,000 exons mapped to 654 genes that were not significantly enriched in known pathways or gene ontology terms (data not shown).

Unsupervised hierarchical clustering formed two major clusters (1 and 2), the second of which was further split into 2 clusters (2A and 2B; **Fig 4.4A**). Cluster 2B contained two final large clusters: 2B1 and 2B2. The corresponding groups of patients had several distinguishing clinical characteristics. First, using Kaplan-Meier curves, I observed a significant difference in survival time (**Fig 4.4B**), with cluster 1 patients having the best prognosis (median survival = 1,024 days), and cluster 2B1 the second-best prognosis (median survival = 551 days). The two poor-prognosis groups (cluster 2A median survival = 447 days, cluster 2B2 median survival = 345 days) were not significantly different from each other, but their median survival was significantly smaller than both good-prognosis curves (log-rank test p-values in **Fig 4.4B**).

Overall survival at the 2-year time point was quite similar for the patients in the poor-prognosis clusters (21.1% for 2A, 15.0% for 2B2, **Fig. 4.5A**), and much lower than the 2-year survival of patients in clusters 1 and 2B1 (69.5% and 39.6%, respectively). However, the conditional 5-year survival rate, which is the probability of survival to five years for those patients alive at two years, distinguished cluster 2B2 from cluster 2A as having a particularly dismal prognosis (5.9% for 2B2, 20.0% for 2A, **Fig 4.5A**).

Age is a known prognostic factor for survival of GBM patients (Ohgaki and Kleihues 2007), so I sought to determine whether this variable differed significantly between the four patient groups. As expected, the patients in the best prognosis group (cluster 1) were the youngest (median age = 33 years; **Fig 4.5A**) by a significant margin compared to the other groups (**Fig 4.5A**; all pairwise Student's T-tests, $P < 6 \times 10^{-3}$). In contrast, the subgroups in cluster 2 did not differ significantly from eachother (i.e. 2A, 2B1, 2B2), indicating that clinical factors other than age may distinguish these patients (**Fig 4.5A**).

4.2.4 Group-specific differences in response to Temozolomide treatment

To determine whether clinically relevant features other than age and overall survival might distinguish patient groups, I investigated survival outcomes based on treatment with Temozolomide. Temozolomide is a recently approved chemotherapeutic drug (Cohen et al. 2005; Stupp et al. 2005) that is rapidly absorbed after oral administration, spontaneously converts to its active form without hepatic metabolism, efficiently penetrates the blood-brain barrier, and has mild and predictable side effects. In a program of concomitant administration with radiotherapy, Temozolomide confers a small yet

significant survival advantage to patients concurrently undergoing radiotherapy (increasing median survival from 12.1 to 14.6 months). Upon its approval, it has become the standard treatment for GBM, and 100 (40.8%) of the 245 TCGA patients had concomitant Temazolomide and radiation as part of their therapy. I therefore sought to determine whether improved survival due to this treatment was evident within each patient cluster (derived in 4.2.3).

I partitioned patients from each group into those treated with Temozolomize and those treated with other chemotherapeutic agents, and then reassessed survival using the Kaplan-Meier method (**Fig 4.5B**). Due to the small number of patients in cluster 1 that were treated with agents other than Temozolomide (n=2), no significant difference was detectable in survival (log-rank test, P = 0.29). Patients with the second-best prognosis (cluster 2B1) had no observable difference in survival outcomes due to treatment (log-rank test, P = 0.96). In contrast, the two poor-prognosis groups differed dramatically in terms of treatment-specific survival, with Temozolomide having no significant effect on cluster 2B2 patients (log-rank test, P = 0.063), but leading to a significant increase in the survival rate of cluster 2A patients (log-rank test, $P = 1.3 \times 10^{-3}$). In the 2A group of patients, median survival increased 1.8-fold in response to treatment with Temozolomide (from 313.0 days to 547.5 days).

4.2.5 Comparison to other GBM clustering methods

A recent analysis of the same TCGA dataset (Verhaak et al. 2010) used unsupervised hierarchical clustering of gene expression values to subtype GBM samples into four groups. These groups, (Proneural, Neural, Classical, and Mesenchymal), were distinguished by a panel of somatic mutation and copy number changes in genes known to belong to pathways relevant to GBM pathogenesis (Verhaak et al. 2010). In addition, each had gene expression signatures reminiscent of mature cell types (neurons, oligodendrocytes, astrocytes, and astroglial cells), and therefore likely represent tumors evolved from distinct cell populations. To determine whether the prognostic value of these gene-expression derived clusters was comparable to that of the exon-expression derived clusters, I compared the survival rate of patients in all groups.

As described in previous sections, the clusters derived from antisense-correlated splicing events represent a range of good to poor prognosis patient groups (**Fig 4.6**). In contrast,

there was no significant difference between the survival outcomes of individuals grouped using gene-expression values (all pairwise student's t-tests with $p > 0.05$; **Fig. 4.6**). Thus, while the gene-expression derived subtypes can indicate the cell types from which tumors evolved, and the likely spectrum of somatic mutations, the groups derived from the exon-expression data can be more readily adapted to use as prognostic tools.

4.2.6 Antisense-correlated alternative splicing of genes with putative driver roles in GBM pathogenesis

Molecular studies of GBM copy number alterations, somatic mutations, DNA methylation patterns, and gene expression changes have revealed that the etiology of GBM is generally based on aberrations in genes belonging to three critical signaling pathways: growth factor receptors (PI3K and Ras), p53, and RB (Phillips et al. 2006; TCGA 2008; Cerami et al. 2010). I compiled a list of 82 genes from these publications and from the Ingenuity Pathway Analysis database (www.ingenuity.com), and asked whether any had known or novel antisense transcription, and whether they passed the expression filters in the GBM dataset. Thirty-three genes (40.2%) met these criteria, and for 19 (57.6%) of these, significant antisense-correlated splicing events were observed (**Table 4.2**). The majority of these genes had isoforms that were either cancer-specific (89.5%) or GBM-specific (68.4%). Interestingly, 10 (76.9%) of the genes with GBM-specific isoforms were members of the PI3K pathway, suggesting that splicing regulation may be particularly relevant to the functioning of this signaling pathway.

Six of the 17 candidate driver genes with cancer-specific isoforms had exons that were part of the list of 1,000 highly variant exons used in section 4.2.4 to distinguish good from poor prognosis patients (EGFR, AVIL, FOXO1, TLC1, PLCL2, PLCB1; **Table 4.2**). Of these, EGFR had the highest number of antisense-correlated alternatively spliced exons (**Fig. 4.7**). In total, 17 EGFR exons were spliced in a manner significantly correlated with the novel antisense construct, and 16 of these were negatively correlated, indicating that in the presence of the antisense transcript they were excluded from one or more of the EGFR isoforms. Seven of these exons had strong correlation values (< -0.6), indicating that the antisense transcript may have a relatively strong role in influencing their exclusion. A subset of these exons encode the tyrosine kinase domain of the gene (**Fig 4.7**), and potentially alters the proportion of EGFR isoforms that contain a complete kinase domain.

Of particular clinical interest are those antisense-correlated splicing events that associate with prognosis. Finding such events can be achieved using the Cox proportional hazards method (Methods). To assess whether the genes previously identified as putative driver genes in GBM had prognostic splicing events, I applied the Cox proportional hazards model to EGFR, AVIL, FOXO1, TLC1, PLCL2, and PLCB1 exons. Fitting the Cox proportional hazards model revealed that one probeset in the PLCL2 gene was associated with survival, and this association remained significant after multiple test correction (corrected P = 0.038). Since this probeset mapped to an intron, it may represent a novel exon of PLCL2. However, the lack of EST-based support for a novel exon increases the likelihood that this probeset may instead represent a retained intron event. Interestingly, inclusion of the region identified by this probeset (i.e. a large SI value) was observed in patients with poor prognosis (median survival = 484 days, n=109) while its exclusion (a small SI value) was observed in patients with a better prognosis (median survival = 682 days, n=136).

4.3 Discussion

Multiple genome and transcriptome-profiling technologies are being leveraged by The Cancer Genome Atlas Research Network to generate massive amounts of genomic profiling data designed to further our understanding of cancer biology (TCGA 2008). The challenge engendered by such data is in devising ways to discern biologically meaningful signal from noise. Successful approaches involve identifying genes expressed or mutated above background (random) levels. I employed a similar strategy to discern biologically relevant information from alternative splicing events. Previous reports have documented the prevalent and tissue specific patterns of splicing events (Wang et al. 2008a), the high levels of alternative splicing events in the brain (Grabowski and Black 2001), and have validated the use of splicing arrays in identifying subtypes of brain cancer (French et al. 2007). In contrast to these studies, I focused my analysis on the subset of alternative splicing events that were correlated to the level of *cis*-encoded antisense transcription. Antisense transcription has been shown to mediate splicing regulation (Hastings et al. 2000), and I have recently shown that antisense-mediated splicing regulation is likely a prevalent phenomenon in the human genome. The biological relevance of antisense-correlated splicing events is evident from the work presented here, since these events are not only prevalent in normal human tissues, but can

also serve to identify sub-groups of normal tissues through unsupervised clustering methods.

Having established the relevance of antisense-correlated splicing events in normal cells, I focused on the potential utility of these signals in uncovering subtypes of GBM, a brain tumor with exceptionally poor prognosis. Gene-expression values have been previously used to segregate GBM patients into subtypes that were generally representative of driver somatic mutations and the cell types of tumor origin (Verhaak et al. 2010). In a similar approach, I used antisense-correlated alternative splicing events to identify four subtypes of GBM. One caveat of this approach is that the robustness of the identified subtypes have to be validated (using bootstrapping methods for instance). In contrast to previous gene-expression derived groups, the subtypes identified using exon-expression information distinguished patients with a range of poor to good survival outcomes. This is the first report using molecular profiling of antisense-correlated alternative splicing events as a prognostic marker in a large cohort of cancer patients.

Of specific clinical relevance was the finding that patients in one of the two poor-prognosis groups had an improved response to concurrent Temozolomide and radiotherapy, which is part of the current standard of care for GBM patients. One explanation for the response of specific patients to Temozolomide could be epigenetic silencing of the DNA-repair gene O^6-methylguanine–DNA methyltransferase (MGMT) (Hegi et al. 2005). A methylated MGMT promoter has been shown to confer sensitivity to drug action and radiation therapy, since promoter methylation decreases gene expression at this locus, and consequently prevents repair of drug-induced DNA lesions (**Fig. 4.8**, (Chakravarti et al. 2006)). However, the expression of MGMT did not contribute to my results since its novel antisense partner did not pass expression thresholds, and consequently the gene was not analyzed. Thus, the demonstrated ability to identify Temozolomide-sensitive patients based on antisense-correlated splicing events indicates that other predictors exist for sensitivity to Temozolomide. This observation is in accord with previous findings that a methylated MGMT promoter only conferred a survival advantage to 62% of patients treated with Temozolomide (Hegi et al. 2005). Genes other than MGMT are therefore likely to be involved in Temozolomide resistance and sensitivity, and importantly, include genes that exhibit antisense-correlated splicing

events specific to cancers. Since there are no known MGMT homologs, these could be genes that regulate MGMT promoter methylation, or genes acting in a closely related pathway to influence dsDNA repair. The antisense-correlated splicing events identified in this chapter essentially represent a short-list of genes from which we can gain a more complete understanding of the molecular basis for Temozolomide sensitivity.

Identifying newly diagnosed patients who are not expected to respond to Temozolomide would be of considerable clinical value, as those patients may derive more benefit from enrolling in clinical trials rather than receiving standard care. One approach to identifying Temozolomide-responsive patients is to first find a subset of exons whose splicing status is significantly associated with the survival of group 2A patients. The Cox proportional hazards model can be used to test such associations. As an example of how this might be done, I carried out a similar analysis of association between survival and the splicing index of six candidate driver genes in GBM. One of 13 probesets thus tested was found to have a significant association with prognosis, such that its inclusion was indicative of poor survival and its exclusion indicative of good survival. Using a similar strategy, a survey of the 1,000 highly variant antisense-correlated splicing events should indentify a subset of exons associated with Temozolomide response. Once a set of predictive exons is found, their sensitivity and specificity can be defined in an unrelated set of GBM patients for which Temozolomide response is known. Ultimately, a minimal set of sensitive and specific splicing events can be developed into a clinical test, for instance a PCR-based assay that detects the presence or absence of the predictive exons.

Antisense-correlated splicing events have been found to occur in nearly half of the candidate GBM driver genes, and specifically in the PI3K signaling pathway. Understanding the consequences of these splicing events on both gene function and the function of the signaling pathway could provide further insight into the etiology of GBMs. For example, an analysis of the exons that encode the EGFR protein kinase domain revealed that antisense-correlated splicing events have the potential to significantly impact protein functionality by generating EGRF isoforms with an incomplete protein kinase domain. This highlights that a good strategy for future identification of putative drug targets may be to focus on the subset of alternatively spliced exons that encode protein domains. This strategy has been previously employed

for the EGFRvIII isoform, which is an EGFR splice variant prevalent in cancers, including GBM. This isoform is missing part of the extracellular domain encoded by exons 2-7, generally due to an in-frame deletion. The fusion point is the target of therapies that rely on monoclonal antibodies to recognize the tumor-specific epitope (Kuan et al. 2001). Thus, one potential application of the methods described in this chapter is the identification of putative novel drug targets.

4.4 Methods

4.4.1 Affymetrix exon array data

A total of 324 GBM and 518 OVC samples were downloaded from the TCGA data portal (tcga-data.nci.nih.gov). After averaging technical replicates, a total of 279 GBM samples remained, of which 10 were normal tissue (epileptic brain samples), and 3 were universal controls. The panel of normal samples included the 10 GBM normal tissues, plus a subset of arrays from the following GEO datasets: lymphoblastoid cell lines (GSE9703), erythrocytes (GSE14588), stomach (GSE13195), thymus (GSE11967), lung (GSE12236), prostate (GSE12378), stem cells (GSE18698), spinal cord (GSE18920), colon (GSE19163), blood (GSE19470), fibroblasts (GSE21440), adult brain (GSE9385), fetal brain (GSE13344), and one Affymetrix-generated tissue panel containing breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testes, and thyroid samples (references in **Table 4.1**). Technical and biological replicates in the set of 304 samples were averaged to yield a final set of 230 samples.

4.4.2 CEL file processing

All Affymetrix CEL files were background corrected and normalized using Affymetrix Power Tools (www.affymetrix.com). Background correction and quantile normalization was computed using RMA-sketch. Resulting probeset expression values were filtered based on variance such that probesets with less than median variance were excluded from further analyses. Variance was calculated across the GBM, OVC, and Normal samples independently.

4.4.3 Assessing antisense-correlation of alternative splicing events

Probesets included in the analysis were those with greater than median variance in a given dataset. Genes included in the analysis were those with at least 20% of probesets

that passed the variance threshold. Probesets were mapped to the sense and antisense strands of genes using custom Perl scripts. As previously described (Morrissy et al. 2010a), a splice index (SI) was calculated for each probeset in the OVC, GBM, or Normal dataset independently, as well as for the whole dataset. Briefly, the SI is the fraction of gene expression that corresponds to one exon, and changes in SI indicate alternative splicing events. Correlations between the SI of each exon in a sense gene and the expression of the corresponding antisense gene were computed in R using the cor.test function (R_Development_Core_Team 2008). Associated correlation p-values (Best and Roberts 1975) were multiple-test corrected using the Bonferroni method (Wright 1992). At novel SAS loci, the expression of probesets mapping within sense gene boundaries but on the opposite strand was averaged and used to approximate antisense expression (refered to as "antisense construct").

4.4.4 Clustering analysis

Exons with antisense-correlated splicing were subjected to unsupervised hierarchical clustering using the TM4 Microarray Software Suite (Saeed et al. 2006). Default parameters were used (average linkage based on Pearson correlations), and clustering was performed using exon SI values.

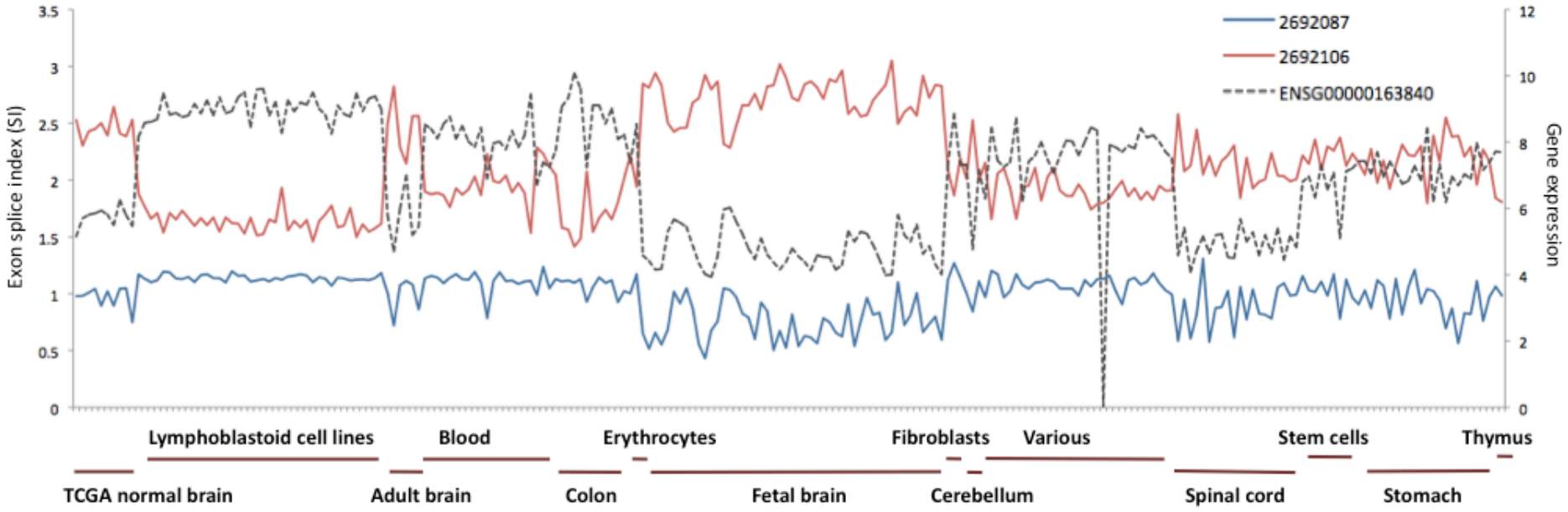
4.4.5 Survival analysis

To test for differences in survival times between the patients in the four GBM clusters, the "survival" package in R was used to generate Kaplan-Meier curves and conduct log-rank tests. Boxplots of survival values were also generated in R. Associations between the splice index values of individual exons and clinical variables (such as survival) were calculated using the "coxph" method in R, and resulting p-values were multiple test corrected by the Benjamini-Hochberg approach in "p.adjust".

Figure 4.1 Antisense-correlated splicing events.

(A) The splice indexes (SI) of two significantly antisense-correlated exons are shown along with the expression of the antisense gene (dotted black line) in the set of normal libraries (x-axis, labels shown for every third library). One exon is negatively correlated (red line, r = -0.92), and one is positively correlated (blue line, r = 0.81). The two probesets map to exons of the gene PARP9, which is encoded antisense to the gene DTX3L (ENSG00000163840). (B) The SI of a sense gene exon at a novel SAS gene locus (red line, r = -0.87) is shown along with the expression of the antisense construct (dotted black line). The sense exon is part of the novel SAS gene CLIP3.

A



B

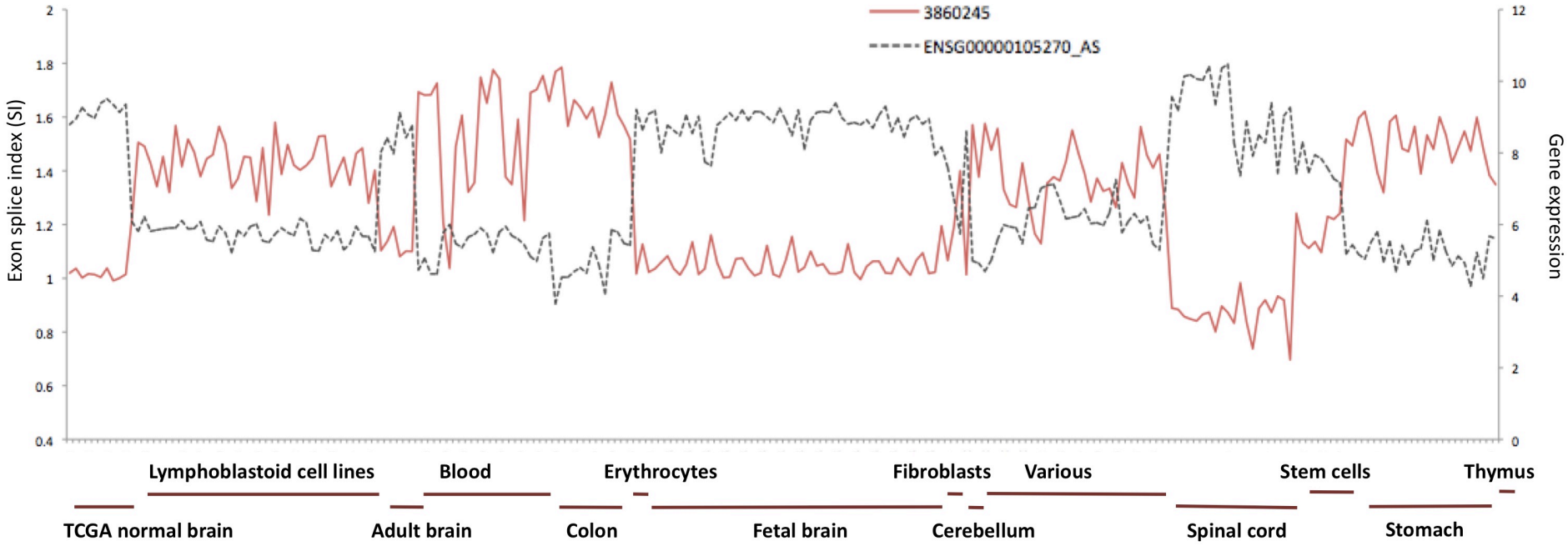

3860245
ENSG00000105270_AS
Exon splice index (SI)
Gene expression
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
12
10
8
6
4
2
0
Lymphoblastoid cell lines
Blood
Erythrocytes
Fibroblasts
Various
Stem cells
Thymus
TCGA normal brain
Adult brain
Colon
Fetal brain
Cerebellum
Spinal cord
Stomach

Figure 4.2 Prevalence of antisense-correlated splicing events.

(A) The number of genes and exons passing expression thresholds in Normal, GBM and OVC samples (see Methods). The number of genes with significant SAS-correlated splicing as well as the number of spliced exons are enumerated. (B) Number of exons with significant antisense-correlated splicing observed in particular subsets of the data (left panel), and the number genes they map to (right panel).

A

	Tissues	Arrays	Expressed SAS genes	Expressed SAS exons	Genes with SAS-correlated splicing	Exons with SAS-correlated splicing
GBM	1	266	4,594	83,646	2,179	9,410
OVC	1	518	4,739	90,287	3,099	14,610
Normals	26	230	4,801	107,179	3,312	17,420

B

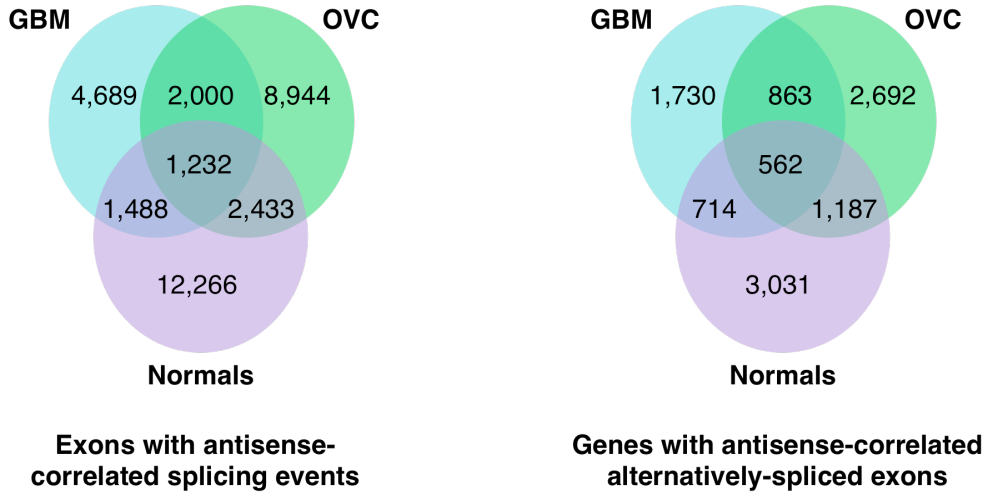


Figure 4.3 Unsupervised hierarchical clustering of exons with antisense-correlated splicing in Normal tissues.

Clusters specific to each input tissue are generated using unsupervised hierarchical clustering. Tissue group bars are displayed above the dendogram, and are annotated in the legend along with the number of samples per group.

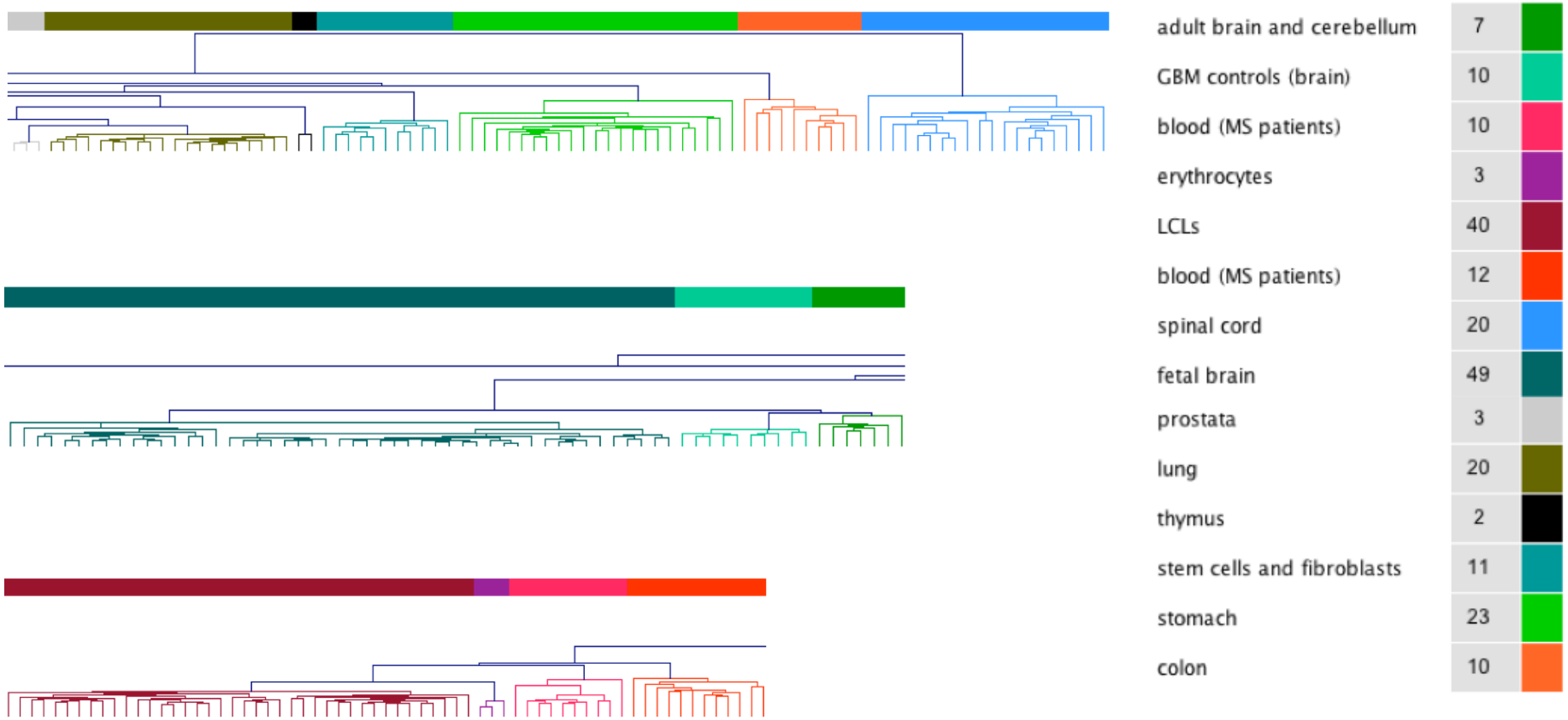
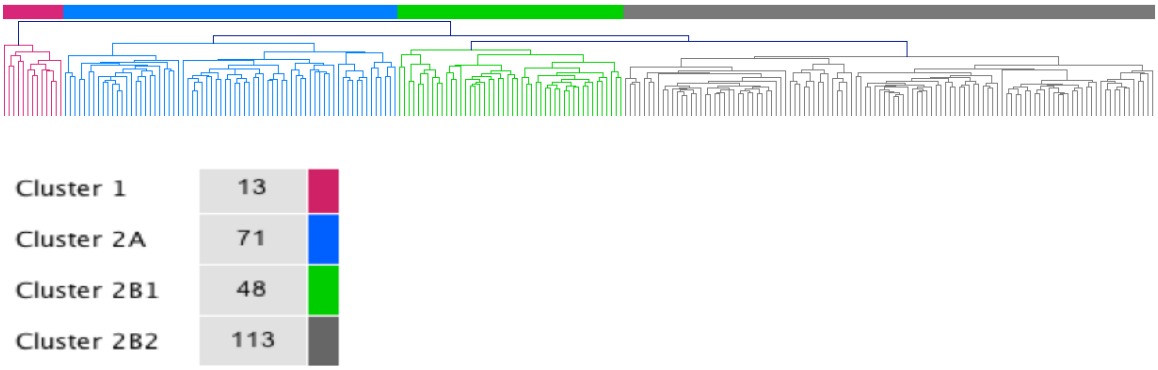


Figure 4.4 GBM subtypes.

(A) Unsupervised hierarchical clustering of the 1,000 most variant exons with antisense-correlated splicing in GBM yields four clusters. Group bars are displayed above the dendogram, and are annotated in the legend along with the number of samples per group. (B). Survival curves corresponding to the four clusters are displayed in a Kaplan-Meier graph. Number of samples (n), and p-values of differences between curves are annotated in the legend.

A



B

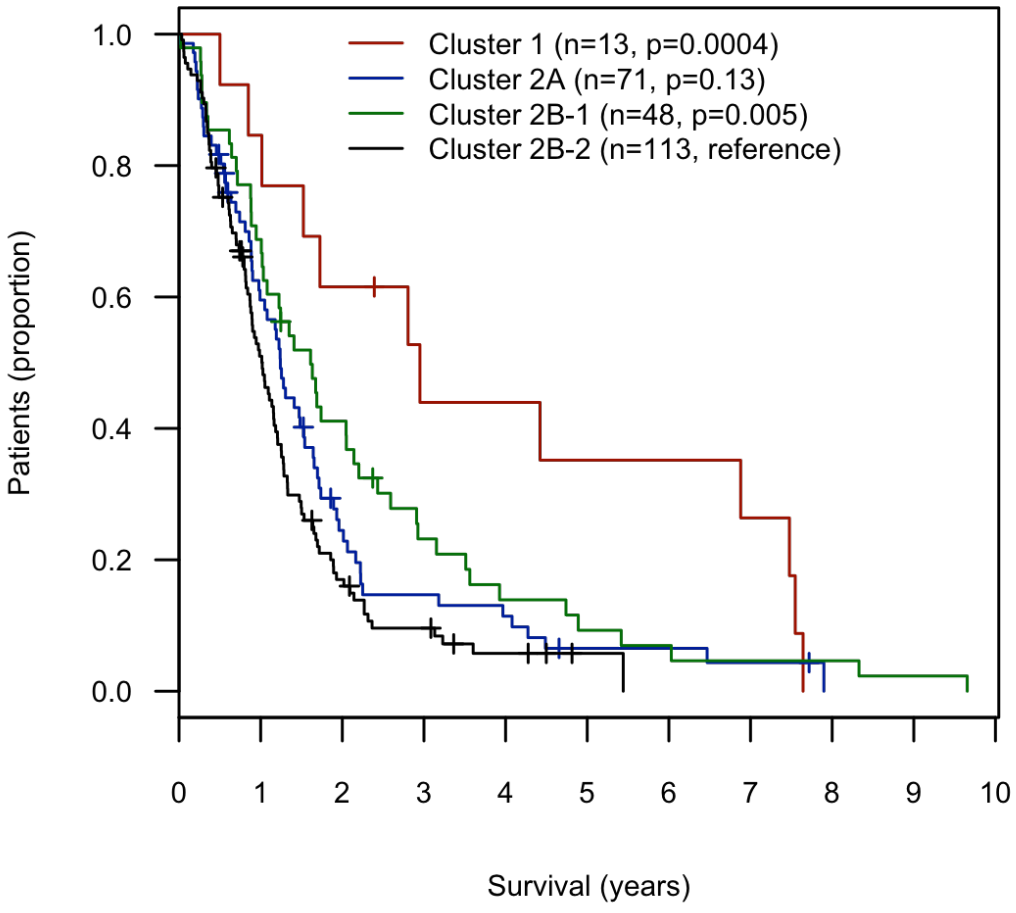


Figure 4.5 Clinical characteristics of patient clusters.

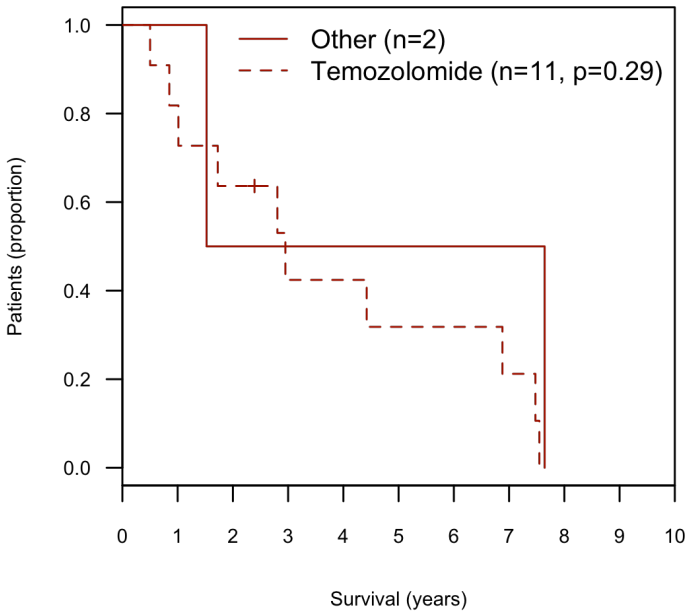
(A) Median survival and age, overall 1-year and 2-year survival, and conditional 5-year survival (B). Kaplan-Meier curves were calculated separately for patients treated with Temozolomide or another chemotherapeutic. P-values of differences between the curves are annotated in the legend, along with the number of samples (n).

A

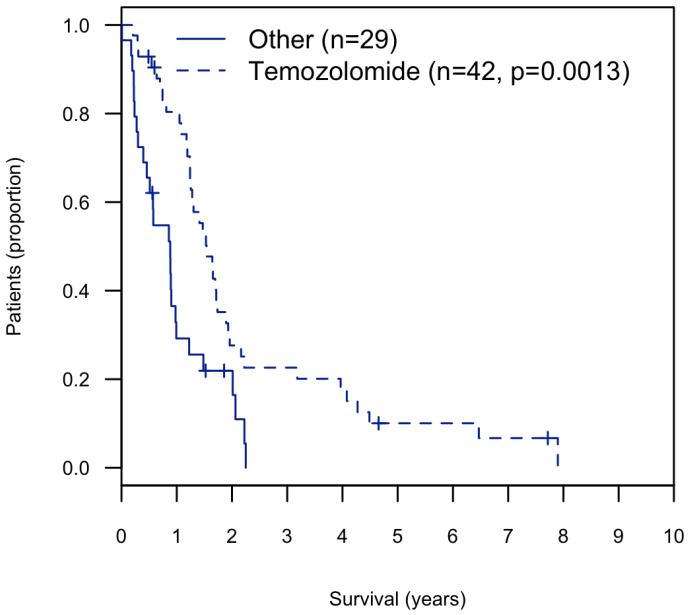
	Number of Patients	Median survival (days)	Median age	1-Year Survival (%)	2-Year Survival (%)	Conditional 5-year Survival *
Cluster 1	13	1,024	33	84.6	61.5	50.0
Cluster 2A	71	447	56	56.3	21.1	20.0
Cluster 2B1	48	551	58.5	68.8	39.6	21.0
Cluster 2B2	113	345	57	47.8	15.0	5.9

* 5-year survival rate (%) was calculated for the subset of patients still alive at 2 years.

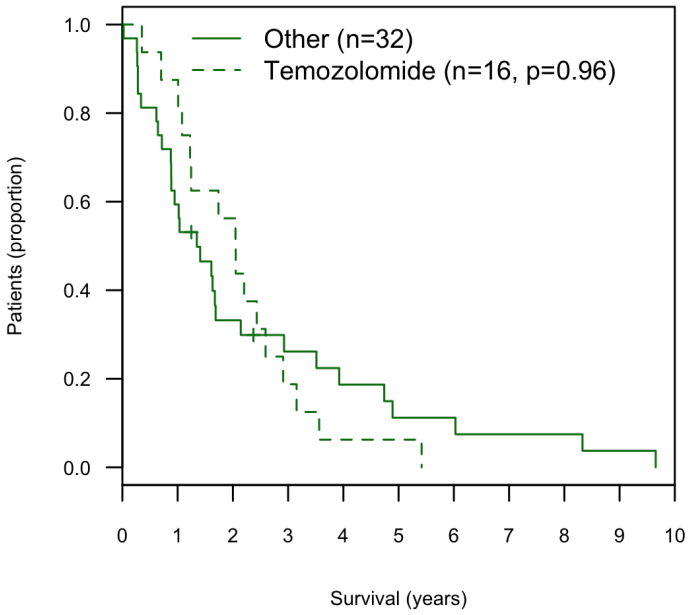
B



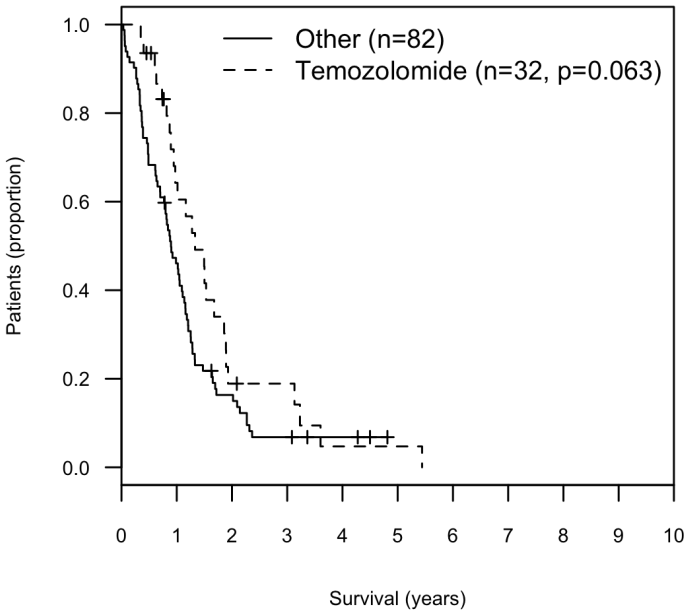
Cluster 1 - Temozolomide vs Other



Cluster 2A - Temozolomide vs Other



Cluster 2B1 - Temozolomide vs Other

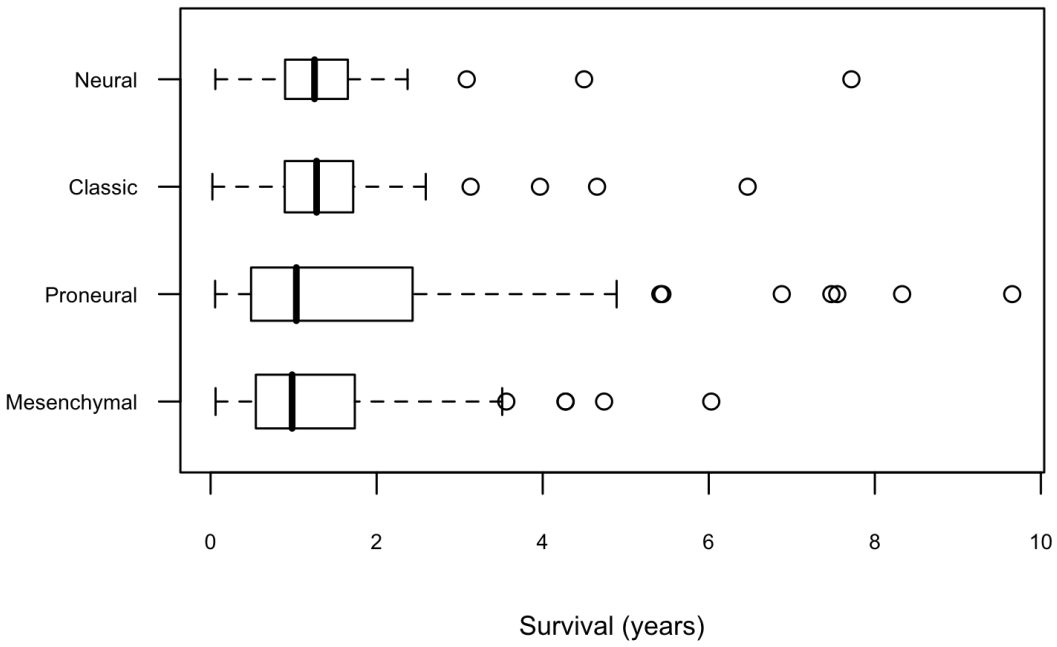


Cluster 2B2 - Temozolomide vs Other

Figure 4.6 GBM subtypes derived from gene-expression versus exon-expression data.

(A) Gene-expression based clusters (Verhaak et al. 2010). (B) Clusters derived from antisense-correlated splicing events. Boxes reflect the IQR; whiskers extend to 1.5 x SD; box heights represent the number of samples in each group (n).

A



B

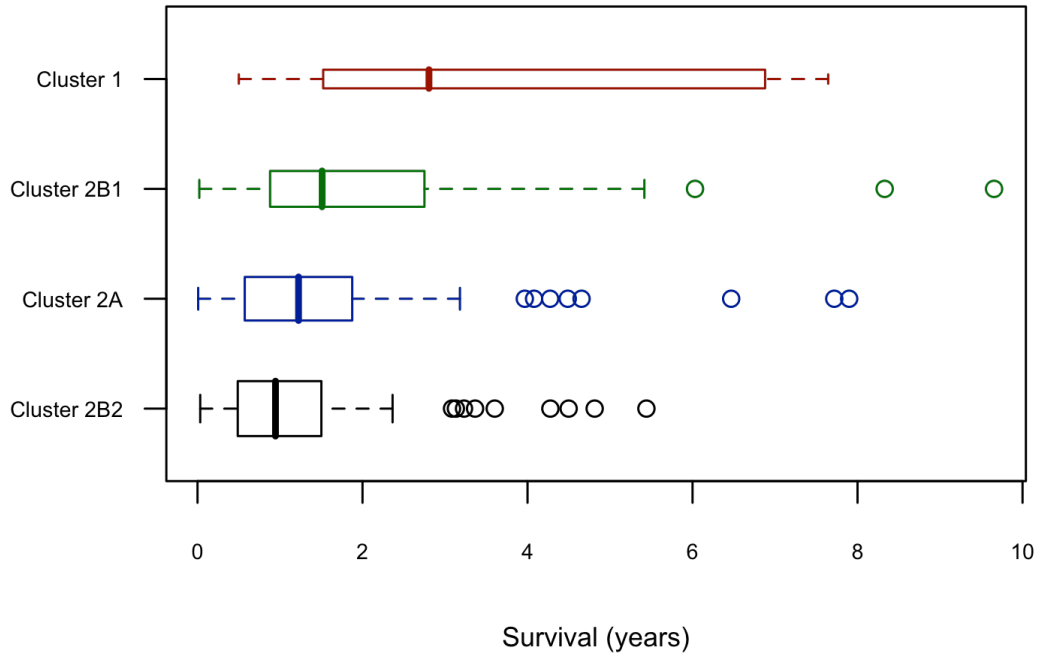


Figure 4.7 Antisense-correlated splicing events in EGFR and PLCL2.

(A) Five isoforms of EGFR are diagramed as rectangles (exons) connected by directional lines (introns). Exonic probesets are shown below the corresponding exons, and the colour indicates the degree of antisense-correlated splicing at each exon. Negative correlations are red bars, positive correlations are green bars, and exons without antisense-correlated splicing events are gray bars. The subset of 7 exons with correlations < -0.6 are shown in the lower track. A horizontal blue line indicates exons 18-25, which encode the kinase domain. (B). A probeset with positive antisense-correlated splicing maps to the intron of PLCL2. Inclusion of the corresponding region in the final mRNA is associated with poor prognosis.

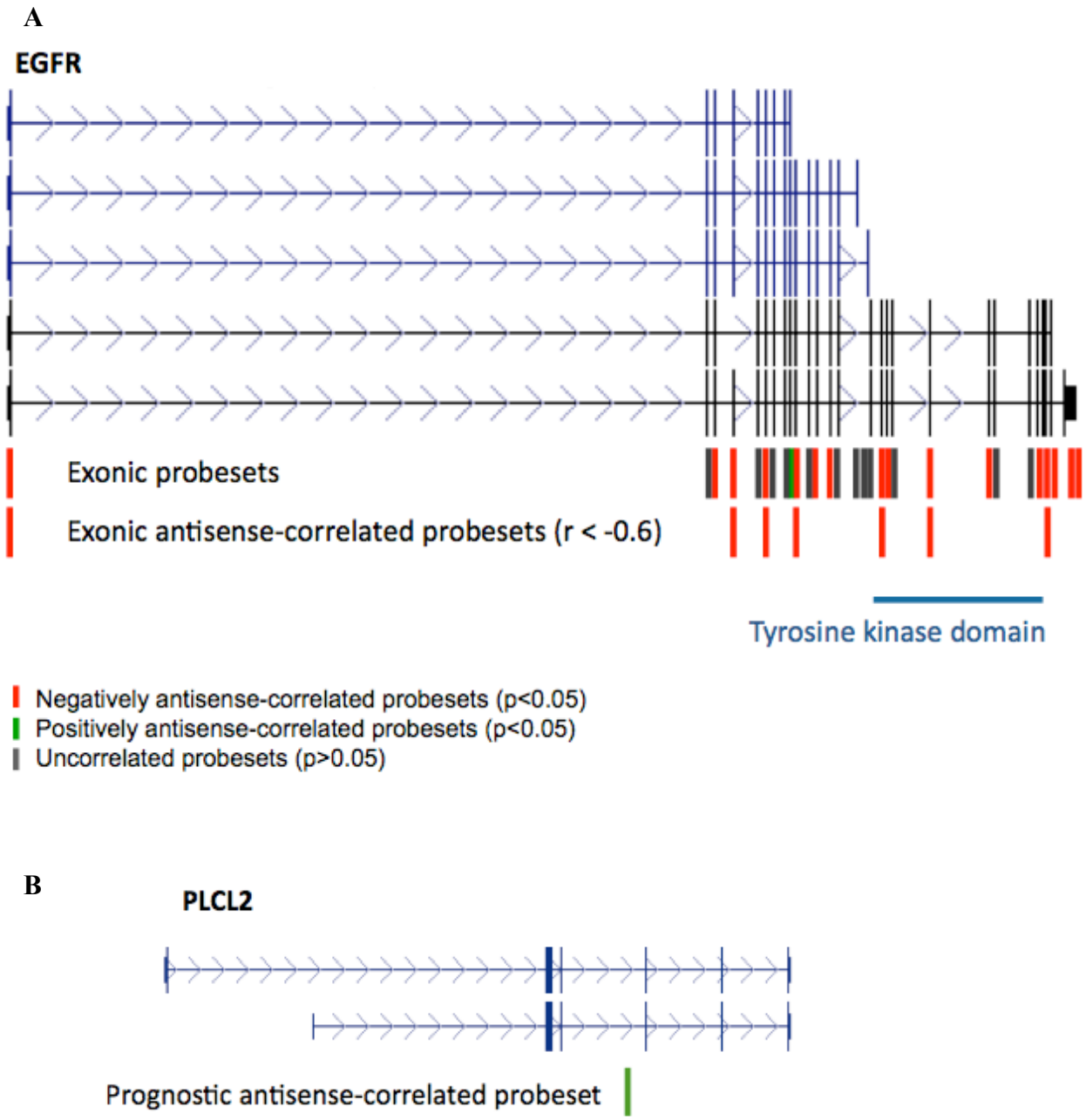


Figure 4.8 Illustration of the role of MGMT in GBM treatment.

A schematic representation of the role of MGMT (dark blue) in responding to Temozolomide-mediated addition of methyl groups (orange) to the O^6 position of guanines. In patients with a silenced MGMT gene (i.e. via a methylated promoter), guanine-methyl groups are not removed, and lead to an increase in the number of dsDNA breaks induced by radiation therapy. In patients with active MGMT, methyl groups are removed, leading to a resistant phenotype. Patients originally sensitive to concomitant treatment with Temozolomide and radiation can become resistant by gaining inactivating mutations in the mismatch repair pathway.

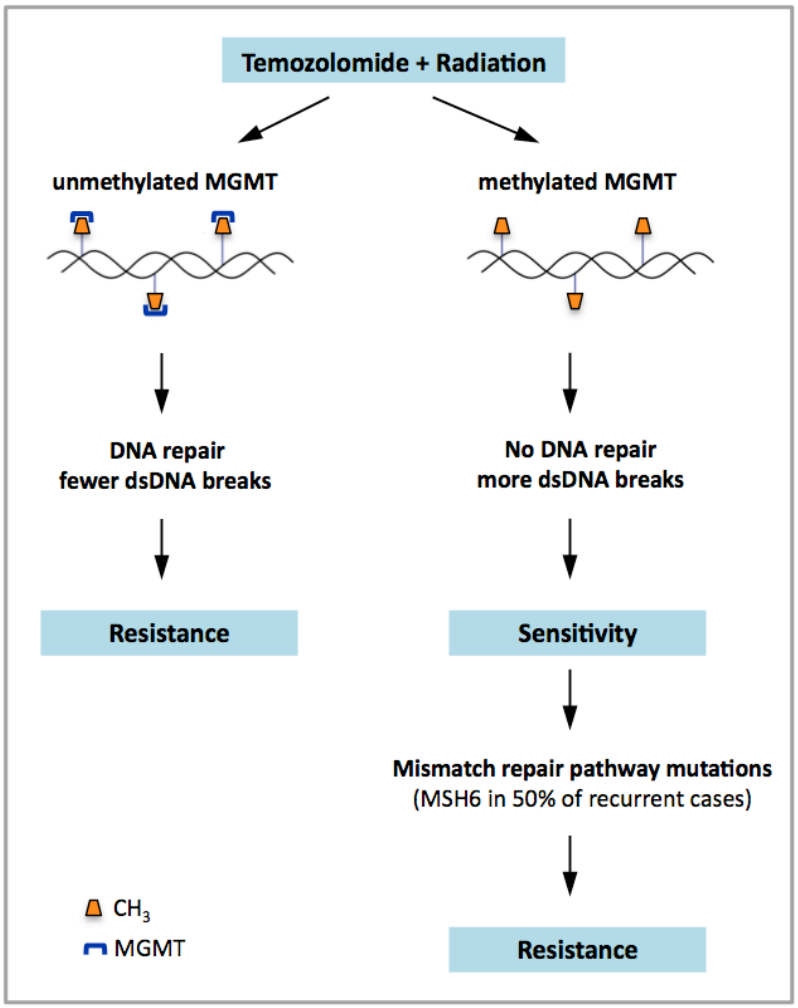


Table 4.1 Summary of Affymetrix exon array data.

Tissue of origin and GEO identifiers of normal and cancer datasets.

GEO Series	Samples	*** Unique	Tissue	Description
GSE13344	94	49	fetal brain	96 samples from left and right hemisphere of 13 brain regions of 2nd trimester fetuses (L and R samples were treated as biological replicates); (Johnson et al. 2009)
GSE9385	6	6	adult brain	6 adult brain samples; (French et al. 2007)
GSE21440	3	3	fibroblasts	3 control fibroblast cultures
GSE19470	22	22	blood	CD3+ cells from whole blood; 2 timepoints of differentiation
GSE19163	12	10	colon	10 cell lines of normal colon, 2 profiled in duplicate; (Mojica and Hawthorn 2010)
GSE18920	20	20	spinal chord	10 lumbar spinal cords, Motor Neuron and Anterior Horn profiled in each individual, (Rabin et al. 2010)
GSE18698	8	8	stem cells	3 neonatal unrestricted somatic stem cells (USSC) from cord blood; 3 bone-marrow derived mesenchymal adult stem cells (BM-MSC) and 2 adipose tissue-derived adult stem cells (AdAS); (Jansen et al. 2010)
GSE12378	3	3	prostate	3 normal prostates; (Jhavar et al. 2009)
GSE12236	20	20	lung	20 normal lung tissue samples adjacent to adenocarcinomas; (Xi et al. 2008)
GSE11967	2	2	thymus	2 normal thymus samples; (Soreq et al. 2008)
GSE13195	23	23	stomach	Gastric mucosal tissue
GSE14588	8	3	erythrocytes	developing erythrocytes: day 7 (3 biological replicates), day 10 (2 biological replicates), day 14 (3 biological replicates); (Yamamoto et al. 2009)
GSE9703	40	40	LCLs	40 HapMap lymphoblastoid cell lines (LCLs); (Zhang et al. 2008)
*	33	11	various	11 tissues (technical triplicates): breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testes, thyroid
**	10	10	brain	GBM control samples (epileptic brain); (TCGA 2008)

* http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx

** GBM control samples were sourced from the TCGA dataset

*** after replicate arrays are averaged

Table 4.2 Antisense-correlated alternative splicing of GBM candidate driver genes.

Eighty-two genes involved in GBM pathogenesis were collected from the literature. Expressed genes with known or novel antisense expression in the GBM samples are listed in rows, while those without antisense transcription or not expressed are denoted by *. Subsets of genes marked with "Y" have significant antisense-correlated splicing events, cancer-specific events or GBM-specific events. Dark red genes are in the set of genes encoding the 1,000 highly variable exons used in the unsupervised clustering analysis.

Gene ID	Ensembl ID	Antisense-correlated splicing	Cancer-specific isoforms	GBM-specific isoforms
A2M	ENSG00000175899	Y	Y	Y
AKT3	ENSG00000117020	Y	Y	Y
AVIL	**ENSG00000135407**	**Y**	**Y**	**Y**
CCND2	ENSG00000118971	Y	Y	Y
CDKN2C	ENSG00000123080	Y	Y	Y
EGFR	**ENSG00000146648**	**Y**	**Y**	**Y**
PIK3R1	ENSG00000145675	Y	Y	Y
PTEN	ENSG00000171862	Y	Y	Y
SPRY2	ENSG00000136158	Y	Y	Y
APC	ENSG00000134982	Y	Y	Y
FOXO1	**ENSG00000150907**	**Y**	**Y**	**Y**
PLCL2	**ENSG00000154822**	**Y**	**Y**	**Y**
TSC1	**ENSG00000165699**	**Y**	**Y**	**Y**
CCND1	ENSG00000110092	Y	Y	
FGFR1	ENSG00000077782	Y	Y	
KLF6	ENSG00000067082	Y	Y	
PLCB1	**ENSG00000182621**	**Y**	**Y**	
EPHA3	ENSG00000044524	Y		
PTPN11	ENSG00000179295	Y		
FGFR2	ENSG00000066468			
IFNW1	ENSG00000091831			
SH3GL2	ENSG00000107295			
CBL	ENSG00000110395			
FOXO3	ENSG00000118689			
PTPRB	ENSG00000127329			
TUBGCP2	ENSG00000130640			
TBP	ENSG00000132561			
PIK3C2B	ENSG00000133056			
TP53	ENSG00000141510			
FRS2	ENSG00000166225			
CRK	ENSG00000167193			
IRS1	ENSG00000169047			
BNC2	ENSG00000173068			

* MAPK1, RHOA, NRAS, ERBB3, LYZ, NUP50, JAK2, CDK6, MET, MPDZ, GAB1, IFNG, NUP107, RBBP5, FGF23, TEK, PIK3CA, KDR, TUBGCP6, MDM2, KRAS, AGAP2, CDK4, PIM1, RB1, ERBB2, IFNAR1, AKT1, TAF1, VLDLR, CDKN2B, CDKN2A, ADAM12, DOCK1, KIT, SNAPC3, IFNB1, DCTN2, SNRPE, THOC4, MAPK11, IFNA2, NF1, IFNA1, MDM4, HLA-DRA, PDGFRA, FGFR1OP, HSPA1A

5. Conclusions and future directions

Comprehensive genome-wide datasets generated in the last decade have been extensively mined in efforts to better annotate transcribed regions, identify regulatory regions, and profile the differential expression of genes between normal and disease states. Insights into the prevalence of antisense transcription resulting from these efforts, coupled with existing knowledge of antisense-mediated regulation of transcript processing at a small number of loci, indicated that antisense-mediated regulation might be more prevalent than previously appreciated. The primary goals of this thesis were therefore to (1) investigate the prevalence of SAS transcription in the human genome, and (2) to investigate the putative role of antisense transcription in the processing of sense gene transcript isoforms.

5.1 Using Tag-seq to annotate the cancer transcriptome

In **Chapter 1**, I reviewed the transcriptional profiling methods used to analyze the transcriptome prior to the work conducted in **Chapter 2**, which specifically focused on the methods employed to detect antisense transcripts. **Chapter 2** described the increased performance in antisense transcript detection achievable using Tag-seq, a method that took advantage of the ultra high-throughput capabilities of the Illumina sequencing platform. The greater profiling depths achieved by Tag-seq compared to LongSAGE not only allowed more sensitive antisense transcript detection, but increased statistical confidence in measuring differential expression between sense and antisense transcripts in cancerous and normal tissues. Consequently, Tag-seq is extremely well suited to identifying infrequently expressed transcripts below the level of detection for methods such as LongSAGE and microarrays, and yields data suitable for conducting differential expression analysis with a high level of confidence. In contrast to other applications of the Illumina platform, such as RNA-seq, Tag-seq allows distinction of the profiled transcripts strand of origin, and consequently enables analysis of SAS transcripts. A substantial benefit to cancer research is the large number of Tag-seq libraries that have already been generated as part of the Cancer Genome Anatomy Project project.

The primary limitation of the Illumina Tag-seq method is the lack of information regarding transcript structure. In contrast, RNA-seq libraries generate reads spanning whole transcripts, and are thus considerably more informative of transcriptional start and

end sites, and alternative splicing events. Two recently developed computational tools have attempted to deduce the strand of origin from RNA-seq reads (Guttman et al. 2010; Trapnell et al. 2010). These tools rely on splice site sequences and read-pair information to assign a subset of reads to the positive or negative strands, thus generating a semi-quantitative measure of SAS transcription. However, significant development of these algorithms will be required in order to close the gap between their inferred estimates of SAS expression and the precise digital counts generated by Tag-seq.

While it is possible to create strand-specific RNA-seq libraries, this is not routinely done and typical RNA-seq libraries are consequently not amenable to studies of SAS transcription. However, strand-specific RNA-seq library construction might become more popular after a recent in-depth comparison of seven methods that can be used to retain strand-of-origin information (Levin et al. 2010). The authors of this study highlight the performance of the dUTP method as particularly superior in terms of strand-specificity, as well as continuity and evenness of transcript coverage. This method involves addition of deoxyuridine triphosphate (dUTP) instead of deoxythymidine triphosphate (dTTP) to mark the second cDNA strand (Parkhomchuk et al. 2009). The marked strand is degraded before the amplification step, leading to specific amplification of the first cDNA strand, and consequently maintaining strand specificity. Generating RNA-seq libraries using the dUTP protocol only requires a small change in the established RNA-seq protocol, but yields critical transcript structure information at both known and novel genes (Parkhomchuk et al. 2009). For this reason, it seems likely that strand-specific RNA-seq libraries will eventually outnumber existing collections of Tag-seq libraries.

5.2 Characterizing an extensive relationship between antisense transcription and alternative splicing

The numerous documented mechanisms of antisense-mediated regulation reveal the plasticity of these (typically) non-coding transcripts in exerting significant effects upon the processing of *cis*-encoded sense transcripts. Given the lack of genome-wide studies of specific mechanisms of antisense-mediated transcriptional regulation, I set out to investigate the effects of antisense transcription on splicing outcomes on a global scale. **Chapter 3** describes the bioinformatic approach I devised to measure correlations between these two phenomena across the transcriptome. This approach revealed a

widespread and significant correlation between antisense transcription and alternative transcript processing at the majority of human SAS loci. Exon expression, nucleosome occupancy, and PolII occupancy data, were analyzed together in order to explore the properties of SAS sequence overlaps that might affect alternative splicing in a local manner. Decreased polymerase speeds were strongly correlated with nucleosome occupancy, as measured at known SAS loci where both transcripts were annotated. As a result, I proposed a speculative mechanism that links decreased polymerase elongation speed over regions of SAS overlap with local increases of alternative splicing events The analyses described in this chapter therefore highlight the utility of multi-tiered data analysis in exploring alternative splicing events in the context of antisense transcription. These findings indicate that a thorough investigation of splicing outcomes is warranted in future studies of known and novel SAS loci.

One caveat of using exon array data is the inability to distinguish exon connectivity. As such, antisense-mediated splicing events of interest derived from microarray-based studies must ultimately be investigated further using technologies capable of detecting exon connectivity (i.e. RT-PCR; strand-specific RNA-seq libraries). The prevalent correlations between antisense transcription and alternative splicing of sense genes is a strong argument for the adoption of strand-specific RNA-seq protocols, as described in section 5.2. In comparison to exon microarrays, strand-specific RNA-seq libraries would provide more accurate measurements of sense and antisense transcript expression, and allow sensitive measurements of changes in exon usage over a greater dynamic range. In addition, exon connectivity could be directly measured using reads spanning splice sites, allowing detection of antisense-correlated isoforms rather than exons, and ultimately, a more complete understanding of splicing regulation.

5.3 Using antisense-correlated splicing events to identify GBM subtypes

In **Chapter 4**, I collected large exon array datasets representing normal and cancerous tissues and determined that highly variable antisense-correlated splicing events could be used to distinguish different normal tissues from each other. Of specific relevance to our understanding of GBM molecular heterogeneity, I found that cancer-specific antisense-correlated splicing events could be used to distinguish clinically relevant subtypes of GBM. In particular, groups of GBM patients with good to poor survival outcomes could

be characterized using a subset of 1,000 antisense-correlated splicing events, and these groups could be further distinguished into subsets of patients who responded well or poorly to Temozolomide therapy. Thus, for the first time, I provide evidence that antisense-correlated exons represent a clinically relevant subset of splicing events in cancer. Using the approaches described here, such events could be readily investigated in other collections of cancer samples profiled using exon arrays (such as the set of ovarian cancers described in **Chapter 4**), or more generally, in any exon-level expression datasets generated in a strand-specific manner.

The findings in **Chapter 4** open up a number of avenues for future research. A first logical step is to reduce the set of 1,000 highly variable alternatively spliced exons used for clustering into a subset of the most informative events. One way in which this can be done is by using supervised methods that enrich for events correlating to the desired clinical feature, such as survival rate or response to Temozolomide. A short list of informative splicing events can then be analyzed in depth to determine whether the corresponding genes have known roles in cancer biology, and if so, whether the splicing events affect resulting protein structure in ways that might influence function. Alternatively, these genes may be novel players in cancer biology, and their correlation to specific clinical features could be used to short-list pathways involved in the processes of interest, and ultimately, may lead to the discovery of new prognostic and therapeutic targets. Therapeutic strategies targeting aberrantly expressed isoforms in human disease have already been undertaken in animal models as well as humans (Webb et al. 1997; Im et al. 1999). Given the cancer-specific inclusion of a particular exon affecting protein structure, antibody-based therapies can be devised to specifically target resultant cancer-specific epitopoes (reviewed in (Mischel and Cloughesy 2006).

5.4 Conclusions

Overall, the research described in this thesis constitutes a step forward in our understanding of the prevalence and role of SAS genes in the human genome. My findings link antisense transcription with the regulation of alternative transcript processing at the majority of expressed SAS loci. This phenomenon is an important aspect of the molecular heterogeneity of human cancers, and opens new avenues of research in efforts to identify novel prognostic and therapeutic markers.

Bibliography

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**(5013): 1651-1656.

Alfano G, Vitiello C, Caccioppoli C, Caramico T, Carola A, Szego MJ, McInnes RR, Auricchio A, Banfi S. 2005. Natural antisense transcripts associated with genes involved in eye development. *Hum Mol Genet* **14**(7): 913-923.

Annilo T, Kepp K, Laan M. 2009. Natural antisense transcript of natriuretic peptide precursor A (NPPA): structural organization and modulation of NPPA expression. *BMC Molecular Biology* **10**(1): 81.

Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr Biol* **11**(13): 1017-1027.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1): 25-29.

Athanasiadis A, Rich A, Maas S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* **2**(12): e391.

Audic S, Claverie JM. 1997. The significance of digital gene expression profiles. *Genome Res* **7**(10): 986-995.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA et al. 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research* **37**(Database issue): D885.

Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* **71**: 817-846.

Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, Bonilla F, de Herreros AG. 2008. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-ìmesenchymal transition. *Genes & Development* **22**(6): 756-769.

Bentley DR. 2006. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**(6): 545-552.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**(5705): 2242-2246.

Best DJ, Roberts DE. 1975. Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **24**(3): 377-379.

Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T et al. 2004. An overview of Ensembl. *Genome Res* **14**(5): 925-928.

Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D et al. 2009. Virtual terminator nucleotides for next-generation DNA sequencing. *Nat Methods* **6**(8): 593-595.

Butte A. 2002. The use and analysis of microarray data. *Nat Rev Drug Discov* **1**(12): 951-960.

Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K et al. 2002. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* **99**(24): 15524-15529.

Carninci P Kasukawa T Katayama S Gough J Frith MC Maeda N Oyama R Ravasi T Lenhard B Wells C et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**(5740): 1559-1563.

Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM. 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics* **40**(12): 1416-1425.

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**(4): 499-509.

Cerami E, Demir E, Schultz N, Taylor BS, Sander C. 2010. Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS ONE* **5**(2): e8918.

Chakravarti A, Erkkinen MG, Nestler U, Stupp R, Mehta M, Aldape K, Gilbert MR, Black PM, Loeffler JS. 2006. Temozolomide-mediated radiation enhancement in glioblastoma: a report on underlying mechanisms. *Clin Cancer Res* **12**(15): 4738-4746.

Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD. 2005. Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet* **21**(6): 326-329.

Chen J, Sun M, Kent W, Huang X, Xie H, Wang W, Zhou G, Shi R, Rowley J. 2004. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res* **32**: 4812-4820.

Chen J, Sun M, Lee S, Zhou G, Rowley JD, Wang SM. 2002. Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc Natl Acad Sci U S A* **99**(19): 12257-12262.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**(5725): 1149-1154.

Chu J, Dolnick BJ. 2002. Natural antisense (rTSalpha) RNA induces site-specific cleavage of thymidylate synthase mRNA. *Biochim Biophys Acta* **1587**(2-3): 183-193.

Cohen MH, Johnson JR, Pazdur R. 2005. Food and Drug Administration Drug approval summary: temozolomide plus radiation therapy for the treatment of newly diagnosed glioblastoma multiforme. *Clin Cancer Res* **11**(19 Pt 1): 6767-6771.

Dahary D, Elroy-Stein O, Sorek R. 2005. Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Research* **15**(3): 364-368.

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. 2003. A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Molecular Cell* **12**(2): 525-532.

de la Mata M, Kornblihtt AR. 2006. RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nature Structural & Molecular Biology* **13**(11): 973-980.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* **4**(9): R60.61-R60.11.

DiGuistini S, Ralph SG, Lim YW, Holt R, Jones S, Bohlmann J, Breuil C. 2007. Generation and annotation of lodgepole pine and oleoresin-induced expressed sequences from the blue-stain fungus Ophiostoma clavigerum, a Mountain Pine Beetle-associated pathogen. *FEMS Microbiol Lett* **267**(2): 151-158.

Ebisuya M, Yamamoto T, Nakajima M, Nishida E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol* **10**(9) 1106-1113.

ENCODE Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**(5696): 636-640.

Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L et al. 2006. Complex Loci in human and mouse genomes. *PLoS genetics* **2**(4): e47.

Eszterhas SK, Bouhassira EE, Martin DI, Fiering S. 2002. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. *Mol Cell Biol* **22**(2): 469-479.

French PJ, Peeters J, Horsman S, Duijm E, Siccama I, van den Bent MJ, Luider TM, Kros JM, van der Spek P, Sillevis Smitt PA. 2007. Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Res* **67**(12): 5635-5642.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nat Rev Cancer* **4**(3): 177-183.

Galburt EA, Grill SW, Wiedmann A, Lubkowska L, Choy J, Nogales E, Kashlev M, Bustamante C. 2007. Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. *Nature* **446**(7137): 820-823.

Garcia-Blanco M, Baraniak A, Lasda E. 2004. Alternative splicing in disease and therapy. *Nature biotechnology* **22**: 535-546.

Gerhard DS Wagner L Feingold EA Shenmen CM Grouse LH Schuler G Klein SL Old S Rasooly R Good P et al. 2004. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* **14**(10B): 2121-2127.

Ghosh T, Soni K, Scaria V, Halimani M, Bhattacharjee C, Pillai B. 2008. MicroRNA-mediated up-regulation of an alternatively polyadenylated variant of the mouse cytoplasmic {beta}-actin gene. *Nucleic Acids Res* **36**(19): 6318-6332.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**(5439): 531-537.

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

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M et al. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**(7117): 330-336.

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou Y-C, Pugh TJ et al. 2010. Alternative expression analysis by RNA sequencing. *Nat Meth* **5**.

Griffith M, Tang MJ, Griffith OL, Morin RD, Chan SY, Asano JK, Zeng T, Flibotte S, Ally A, Baross A et al. 2008. ALEXA: a microarray design platform for alternative expression analysis. *Nat Meth* **5**(2): 118.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**(1): 91-105.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**(5): 503-510.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100**(1): 57-70.

Hanriot L, Keime C, Gay N, Faure C, Dossat C, Wincker P, Scote-Blachon C, Peyron C, Gandrillon O. 2008. A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome. *BMC Genomics* **9**: 418.

Hastings M, Milcarek C, Martincic K, Peterson M, Munroe S. 1997. Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels. *Nucleic Acids Research* **25**(21): 4296.

Hastings ML, Ingle HA, Lazar MA, Munroe SH. 2000. Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense RNA. *J Biol Chem* **275**(15): 11507-11513.

Haussecker D, Proudfoot NJ. 2005. Dicer-dependent turnover of intergenic transcripts from the human beta-globin gene cluster. *Mol Cell Biol* **25**(21): 9724-9733.

Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, Kros JM, Hainfellner JA, Mason W, Mariani L et al. 2005. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* **352**(10): 997-1003.

Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6**(9): 807-828.

Hirst M, Delaney A, Rogers SA, Schnerch A, Persaud DR, O'Connor MD, Zeng T, Moksa M, Fichter K, Mah D et al. 2007. LongSAGE profiling of nine human embryonic stem cell lines. *Genome Biol* **8**(6): R113.

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

Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ et al. 2007. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci USA* **104**(23): 9758-9763.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T et al. 2002. The Ensembl genome database project. *Nucleic Acids Res* **30**: 38-41.

Hubbell E, Liu W-M, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* **18**(12): 1585-1592.

Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, Chakraverty S, Isaacs A, Grover A et al. 1998. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393**(6686): 702-705.

Im SA, Gomez-Manzano C, Fueyo J, Liu TJ, Ke LD, Kim JS, Lee HY, Steck PA, Kyritsis AP, Yung WK. 1999. Antiangiogenesis treatment for gliomas: transfer of antisense-vascular endothelial growth factor inhibits tumor growth in vivo. *Cancer Res* **59**(4): 895-900.

Imamura T, Yamamoto S, Ohgane J, Hattori N, Tanaka S, Shiota K. 2004. Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochem Biophys Res Commun* **322**(2): 593-600.

Jackson DA, Pombo A, Iborra F. 2000. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *FASEB J* **14**(2): 242-254.

Jansen BJ, Gilissen C, Roelofs H, Schaap-Oziemlak A, Veltman JA, Raymakers RA, Jansen JH, Kogler G, Figdor CG, Torensma R et al. 2010. Functional differences between mesenchymal stem cell populations are reflected by their transcriptome. *Stem Cells Dev* **19**(4): 481-490.

Jhavar S, Brewer D, Edwards S, Kote-Jarai Z, Attard G, Clark J, Flohr P, Christmas T, Thompson A, Parker M et al. 2009. Integration of ERG gene mapping and gene-expression profiling identifies distinct categories of human prostate cancer. *BJU Int* **103**(9): 1256-1269.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**(3): 161-172.

Jiang L, Gonda TA, Gamble MV, Salas M, Seshan V, Tu S, Twaddell WS, Hegyi P, Lazar G, Steele I et al. 2008. Global hypomethylation of genomic DNA in cancer-associated myofibroblasts. *Cancer Res* **68**(23): 9900-9908.

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

Johnson JM, Edwards S, Shoemaker D, Schadt EE. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21**(2): 93-102.

Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanovic D, Geschwind DH, Mane SM, State MW, Sestan N. 2009. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**(4): 494-509.

Johnston PG, Fisher ER, Rockette HE, Fisher B, Wolmark N, Drake JC, Chabner BA, Allegra CJ. 1994. The role of thymidylate synthase expression in prognosis and outcome of adjuvant chemotherapy in patients with rectal cancer. *J Clin Oncol* **12**(12): 2640-2647.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* **36**(16): 5221-5231.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**(3): 331-342.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**(5569): 916-919.

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Research* **15**(7): 987-997.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**(4): 656-664.

Khattra J, Delaney AD, Zhao Y, Siddiqui A, Asano J, McDonald H, Pandoh P, Dhalla N, Prabhu AL, Ma K et al. 2007. Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome Res* **17**(1): 108-116.

Kim VN. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* **6**(5): 376-385.

Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K. 2005. Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Research* **15**: 463-474.

Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res* **13**: 1324-1334.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M et al. 2006. CAGE: cap analysis of gene expression. *Nat Methods* **3**(3): 211-222.

Krystal GW, Armstrong BC, Battey JF. 1990. N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts. *Molecular and Cellular Biology* **10**(8): 4180-4191.

Kuan CT, Wikstrand CJ, Bigner DD. 2001. EGF mutant receptor vIII as a molecular target in cancer therapy. *Endocr Relat Cancer* **8**(2): 83-96.

Kuersten S, Goodwin EB. 2003. The power of the 3' UTR: translational control and development. *Nat Rev Genet* **4**(8): 626-637.

Kulaeva OI, Gaykalova DA, Pestov NA, Golovastov VV, Vassylyev DG, Artsimovitch I, Studitsky VM. 2009. Mechanism of chromatin remodeling and recovery during passage of RNA polymerase II. *Nat Struct Mol Biol* **16**(12): 1272-1278.

Kumar M, Carmichael GG. 1998. Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol Mol Biol Rev* **62**(4): 1415-1434.

Ladd PD, Smith LE, Rabaia NA, Moore JM, Georges SA, Hansen RS, Hagerman RJ, Tassone F, Tapscott SJ, Filippova GN. 2007. An antisense transcript spanning the CGG repeat region of FMR1 is upregulated in premutation carriers but silenced in full mutation individuals. *Hum Mol Genet* **16**(24): 3174-3187.

Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K et al. 1999. A public database for gene expression in human cancers. *Cancer Res* **59**(21): 5403-5407.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, Casari G. 2004. In search of antisense. *Trends in Biochemical Sciences* **29**(2): 88-94.

Lee JT, Davidow LS, Warshawsky D. 1999. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* **21**(4): 400-404.

Lehner B, Williams G, Campbell RD, Sanderson CM. 2002. Antisense transcripts in the human genome. *Trends Genet* **18**(2): 63-65.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**(9): 709-715.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**(7218): 66-72.

Li A, Walling J, Ahn S, Kotliarov Y, Su Q, Quezado M, Oberholtzer JC, Park J, Zenklusen JC, Fine HA. 2009. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Res* **69**(5): 2091-2099.

Licatalosi DD, Darnell RB. 2010. RNA processing and its regulation: global insights into biological networks. *Nature Reviews Genetics* **11**(1): 75-87.

Listerman I, Sapra AK, Neugebauer KM. 2006. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nature Structural & Molecular Biology* **13**(9): 815-822.

Liu G, Loraine A, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose M. 2003. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* **31**: 82-86.

Louro R, Nakaya H, Amaral P, Festa F, Sogayar M, da Silva A, Verjovski-Almeida S, Reis E. 2007. Androgen responsive intronic non-coding RNAs. *BMC Biology* **5**(1): 4.

Lubitz CC, Ugras SK, Kazam JJ, Zhu B, Scognamiglio T, Chen YT, Fahey TJ, 3rd. 2006. Microarray analysis of thyroid nodule fine-needle aspirates accurately classifies benign and malignant lesions. *J Mol Diagn* **8**(4): 490-498.

Malik K, Salpekar A, Hancock A, Moorwood K, Jackson S, Charles A, Brown KW. 2000. Identification of differential methylation of the WT1 antisense regulatory region and relaxation of imprinting in Wilms' tumor. *Cancer Res* **60**(9): 2356-2360.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**(9): 1509-1517.

Matsumura H, Ito A, Saitoh H, Winter P, Kahl G, Reuter M, Kruger DH, Terauchi R. 2005. SuperSAGE. *Cell Microbiol* **7**(1): 11-18.

Mattick J. 2004. RNA regulation: a new genetics? *Nat Rev Genet* **5**: 316-323.

Meegan JM, Marcus PI. 1989. Double-stranded ribonuclease coinduced with interferon. *Science* **244**(4908): 1089-1091.

Michael A, Catherine AB, Judith AB, David B, Heather B, Cherry JM, Allan PD, Kara D, Selina SD, Janan TE et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1): 25.

Mihalich A, Reina M, Mangioni S, Ponti E, Alberti L, Vigano P, Vignali M, Di Blasio AM. 2003. Different Basic Fibroblast Growth Factor and Fibroblast Growth Factor-Antisense Expression in Eutopic Endometrial Stromal Cells Derived from Women with and without Endometriosis. *Journal of Clinical Endocrinology Metabolism* **88**(6): 2853-2859.

Mischel PS, Cloughesy T. 2006. Using molecular information to guide brain tumor therapy. *Nat Clin Pract Neurol* **2**(5): 232-233.

Mojica W, Hawthorn L. 2010. Normal colon epithelium: a dataset for the analysis of gene expression and alternative splicing events in colon disease. *BMC Genomics* **11**: 5.

Moncke-Buchner E, Rothenberg M, Reich S, Wagenfuhr K, Matsumura H, Terauchi R, Kruger DH, Reuter M. 2009. Functional characterization and modulation of the DNA cleavage efficiency of type III restriction endonuclease EcoP15I in its interaction with two sites in the DNA target. *J Mol Biol* **387**(5): 1309-1319.

Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M et al. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* **18**(4): 610-621.

Morrissy AS, Griffith M, Marra M. 2010a. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Mauscript submitted*.

Morrissy AS, Morin RD, Delaney A, Zeng T, McDonald H, Jones S, Zhao Y, Hirst M, Marra MA. 2009. Next-generation tag sequencing for cancer gene expression profiling. *Genome Res* **19**(10): 1825-1835.

Morrissy S, Zhao Y, Delaney A, Asano J, Dhalla N, Li I, McDonald H, Pandoh P, Prabhu AL, Tam A et al. 2010b. Digital gene expression by tag sequencing on the illumina genome analyzer. *Curr Protoc Hum Genet* **Chapter 11**: Unit 11 11 11-36.

Nag A, Narsinh K, Kazerouninia A, Martinson HG. 2006. The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. *RNA* **12**(8): 1534-1544.

Nahkuri S, Taft RJ, Mattick JS. 2009. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle* **8**(20): 3420-3424.

Navarro P, Page DR, Avner P, Rougeulle C. 2006. Tsix-mediated epigenetic switch of a CTCF-flanked region of the Xist promoter determines the Xist transcription program. *Genes Dev* **20**(20): 2787-2792.

Ng P, Tan JJ, Ooi HS, Lee YL, Chiu KP, Fullwood MJ, Srinivasan KG, Perbost C, Du L, Sung WK et al. 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res* **34**(12): e84.

Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH et al. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* **2**(2): 105-111.

Nielsen KL, Hogh AL, Emmersen J. 2006. DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res* **34**(19): e133.

Nikiforova MN, Tseng GC, Steward D, Diorio D, Nikiforov YE. 2008. MicroRNA expression profiling of thyroid tumors: biological significance and diagnostic utility. *J Clin Endocrinol Metab* **93**(5): 1600-1608.

Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J et al. 2002. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res* **12**(11): 1749-1755.

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

Okazaki Y Furuno M Kasukawa T Adachi J Bono H Kondo S Nikaido I Osato N Saito R Suzuki H et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**(6915): 563-573.

Ota T Suzuki Y Nishikawa T Otsuki T Sugiyama T Irie R Wakamatsu A Hayashi K Sato H Nagai K et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36**(1): 40-45.

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37**(18): e123.

Pekarsky Y, Santanam U, Cimmino A, Palamarchuk A, Efanov A, Maximov V, Volinia S, Alder H, Liu CG, Rassenti L et al. 2006. Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. *Cancer Res* **66**(24): 11590-11593.

Peters DG, Kassam AB, Yonas H, O'Hare EH, Ferrell RE, Brufsky AM. 1999. Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. *Nucleic Acids Res* **27**(24): e39.

Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, Misra A, Nigro JM, Colman H, Soroceanu L et al. 2006. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**(3): 157-173.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**(Database issue): D61-65.

Quere R, Manchon L, Lejeune M, Clement O, Pierrat F, Bonafoux B, Commes T, Piquemal D, Marti J. 2004. Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. *Nucleic Acids Res* **32**(20): e163.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **Vienna, Austria**(ISBN 3-900051-07-0): URL http://www.R-project.org.

Rabin SJ, Kim JMH, Baughn M, Libby RT, Kim YJ, Fan Y, Libby RT, La Spada A, Stone B, Ravits J. 2010. Sporadic ALS has compartment-specific aberrant exon splicing and altered cell-matrix adhesion biology. *Hum Mol Genet* **19**(2): 313-328.

Reis E, Nakaya H, Louro R, Canavez F, Flatschart A, Almeida G, Egidio C, Paquola A, Machado A, Festa F et al. 2004. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* **23**(39): 6684-6692.

Riken Genome Exploration Research Group, Genome Science Group, The Fantom Consortium, Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H et al. 2005. Antisense Transcription in the Mammalian Transcriptome. *Science* **309**(5740): 1564.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M et al. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev* **17**(4): 529-540.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**(7): 1311-1323.

Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS et al. 2007. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Research* **38**(Database issue): D620.

Rosenkranz R, Borodina T, Lehrach H, Himmelbauer H. 2008. Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics* **92**(4): 187-194.

Rossignol F, Vache C, Clottes E. 2002. Natural antisense transcripts of hypoxia-inducible factor 1alpha are detected in different normal and tumour human tissues. *Gene* **299**(1-2): 135-140.

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. 2006. TM4 microarray software suite. *Methods Enzymol* **411**: 134-193.

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. 2002. Using the transcriptome to annotate the genome. *Nat Biotechnol* **20**(5): 508-512.

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**(5): 887-898.

Schulze A, Downward J. 2001. Navigating gene expression using microarrays--a technology review. *Nat Cell Biol* **3**(8): E190-195.

Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology* **16**(9): 990-995.

Shearwin KE, Callen BP, Egan JB. 2005. Transcriptional interference--a crash course. *Trends Genet* **21**(6): 339-345.

Shendure J, Church GM. 2002. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol* **3**(9): RESEARCH0044.

Shinagawa T, Ishii S. 2003. Generation of Ski-knockdown mice by expressing a long double-strand RNA from an RNA polymerase II promoter. *Genes Dev* **17**(11): 1340-1345.

Shirasawa S, Harada H, Furugaki K, Akamizu T, Ishikawa N, Ito K, Tamai H, Kuma K, Kubota S, Hiratani H et al. 2004. SNPs in the promoter of a B cell-specific antisense transcript, SAS-ZFAT, determine susceptibility to autoimmune thyroid disease. *Hum Mol Genet* **13**(19): 2221-2231.

Siddiqui AS, Delaney AD, Schnerch A, Griffith OL, Jones SJ, Marra MA. 2006. Sequence biases in large scale gene expression profiling data. *Nucleic Acids Res* **34**(12): e83.

Siddiqui AS, Khattra J, Delaney AD, Zhao Y, Astell C, Asano J, Babakaiff R, Barber S, Beland J, Bohacec S et al. 2005. A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc Natl Acad Sci U S A* **102**(51): 18485-18490.

Silber J, Lim DA, Petritsch C, Persson AI, Maunakea AK, Yu M, Vandenberg SR, Ginzinger DG, James CD, Costello JF et al. 2008. miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. *BMC Med* **6**: 14.

Smilinich NJ, Day CD, Fitzpatrick GV, Caldwell GM, Lossie AC, Cooper PR, Smallwood AC, Joyce JA, Schofield PN, Reik W et al. 1999. A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proc Natl Acad Sci U S A* **96**(14): 8064-8069.

Soreq L, Gilboa-Geffen A, Berrih-Aknin S, Lacoste P, Darvasi A, Soreq E, Bergman H, Soreq H. 2008. Identifying alternative hyper-splicing signatures in MG-thymoma by exon arrays. *PLoS One* **3**(6): e2392.

Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. 2007. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet* **39**(2): 226-231.

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. *Molecular Cell* **36**(2): 245-254.

Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. 2007. Gene-Expression Variation Within and Among Human Populations. *The American Journal of Human Genetics* **80**(3): 502-509.

Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U et al. 2005. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med* **352**(10): 987-996.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**(7): 4465-4470.

Sun M, Hurst LD, Carmichael GG, Chen J. 2006. Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity. *Genome Res* **16**(7): 922-933.

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

Thiriet C, Hayes JJ. 2005. Replication-independent core histone dynamics at transcriptionally active loci in vivo. *Genes Dev* **19**(6): 677-682.

Thrash-Bingham CA, Tartof KD. 1999. aHIF: a natural antisense transcript overexpressed in human renal cancer and during hypoxia. *J Natl Cancer Inst* **91**(2): 143-151.

Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**(9): 996-1001.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5): 511-515.

Tufarelli C, Sloane Stanley J, Garrick D, Sharpe J, Ayyub H, Wood W, Higgs D. 2003. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genetics* **34**(2): 157.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**(7): 1051-1063.

Vanhee-Brossollet C, Vaquero C. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211**(1): 1-9.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* **270**(5235): 484-487.

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

Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP et al. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**(1): 98-110.

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA. 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**(5588): 1833-1837.

von Bubnoff A. 2008. Next-generation sequencing: the race is on. *Cell* **132**(5): 721-723.

Wahl MB, Heinzmann U, Imai K. 2005. LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. *Bioinformatics* **21**(8): 1393-1400.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008a. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(7221): 470-476.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y et al. 2008b. The diploid genome sequence of an Asian individual. *Nature* **456**(7218): 60-65.

Wang Y, Lee AT, Ma JZ, Wang J, Ren J, Yang Y, Tantoso E, Li KB, Ooi LL, Tan P et al. 2008c. Profiling microRNA expression in hepatocellular carcinoma reveals microRNA-224 up-regulation and apoptosis inhibitor-5 as a microRNA-224-specific target. *J Biol Chem* **283**(19): 13205-13215.

Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**(5): 802-813.

Webb A, Cunningham D, Cotter F, Clarke PA, di Stefano F, Ross P, Corbo M, Dziewanowska Z. 1997. BCL-2 antisense therapy in patients with non-Hodgkin lymphoma. *Lancet* **349**(9059): 1137-1141.

Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol* **144**(1): 32-42.

Wederell ED, Bilenky M, Cullum R, Thiessen N, Dagpinar M, Delaney A, Varhol R, Zhao Y, Zeng T, Bernier B et al. 2008. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res* **36**(14): 4549-4564.

Wilusz CJ, Wormington M, Peltz SW. 2001. The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol* **2**(4): 237-246.

Wright SP. 1992. Adjusted P-Values for Simultaneous Inference. *Biometrics* **48**(4): 1005-1013.

Xi L, Feber A, Gupta V, Wu M, Bergemann AD, Landreneau RJ, Litle VR, Pennathur A, Luketich JD, Godfrey TE. 2008. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Res* **36**(20): 6535-6547.

Yamamoto ML, Clark TA, Gee SL, Kang JA, Schweitzer AC, Wickrema A, Conboy JG. 2009. Alternative pre-mRNA splicing switches modulate gene expression in late erythropoiesis. *Blood* **113**(14): 3363-3370.

Yan M-D, Hong C-C, Lai G-M, Cheng A-L, Lin Y-W, Chuang S-E. 2005. Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas. *Human Molecular Genetics* **14**(11): 1465-1474.

Ye BH, Cattoretti G, Shen Q, Zhang J, Hawe N, de Waard R, Leung C, Nouri-Shirazi M, Orazi A, Chaganti RS et al. 1997. The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nat Genet* **16**(2): 161-170.

Yelin R, Dahary D, Sorek R, Levanon E, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol* **21**: 379-386.

Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, Cui H. 2008. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451**(7175): 202-206.

Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ et al. 2008. Evaluation of Genetic Variation Contributing to Differences in Gene Expression between Populations. *The American Journal of Human Genetics* **82**(3): 631-640.

Zhang Y, Liu XS, Liu QR, Wei L. 2006. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Research* **34**(12): 3465-3475.

Appendices

Appendix A CGAP libraries

The 35 Tag-seq and 77 LongSAGE CGAP libraries analyzed in Chapter 2 were grouped by tissue.

Group	Stage	Description	Library	Protocol
bladder	TaG2	Tumor (f)	hs0240	Tag-seq
	T4aG3	Tumor (m)	hs0265	Tag-seq
	T1G3	Tumor (f)	hs0264	Tag-seq
	T1G3	Tumor (f)	hs0241	Tag-seq
	TaG2	Tumor (m)	hs0266	Tag-seq
	NA	Normal urothelium	hs0239	Tag-seq
skin	Stage4	Metastatic melanoma (m)	hs0275	Tag-seq
	NA	Squamous cell carcinoma (f)	hs0284	Tag-seq
	NA	Seborrheic Keratosis (f)	hs0282	Tag-seq
	NA	Seborrheic Keratosis (m)	hs0281	Tag-seq
	NA	Basal cell carcinoma (m)	hs0279	Tag-seq
	NA	Basal cell carcinoma (f)	hs0280	Tag-seq
	NA	Squamous cell carcinoma (m)	hs0283	Tag-seq
	NA	Normal nevus (f)	hs0278	Tag-seq
	Stage4	Metastatic melanoma (f)	hs0276	Tag-seq
	NA	Normal nevus (m)	hs0277	Tag-seq
	Stage2	Primary melanoma (m; nonpigmented)	hs0274	Tag-seq
	Stage2	Primary melanoma (m; pigmented)	hs0273	Tag-seq
	NA	Normal skin (f)	hs0272	Tag-seq
	9months	Foreskin (m)	hs0305	Tag-seq
	NA	Normal skin (m)	hs0271	Tag-seq
uterus	NA	Ovary	hs0194	Tag-seq
	NA	Fallopian tube	hs0195	Tag-seq
bone marrow	NA	Peripheral blood from AML patient (m)	hs0430	Tag-seq
	NA	Peripheral blood from AML patient (m)	hs0429	Tag-seq
	NA	AML	1483	LongSAGE
	NA	AML	1865	LongSAGE
embryonic	NA	H1	1387	LongSAGE
	NA	H1	1390	LongSAGE
	NA	BG01	1603	LongSAGE
	NA	Normal undifferentiated stem cells	hs0238	Tag-seq
	NA	H14	1314	LongSAGE
	NA	H13	1385	LongSAGE
	NA	Hydatidiform mole	hs0324	Tag-seq
	NA	H7	1313	LongSAGE
	NA	HSF6	1311	LongSAGE
	NA	HES4	1383	LongSAGE
	NA	HES3	1312	LongSAGE
	NA	H9	843	LongSAGE
lymph nodes	NA	Burkitt lymphoma	hs0196	Tag-seq
	NA	Diffuse Large B cell lymphoma (m)	hs0204	Tag-seq
	NA	B-cell Malignant Lymphoma	1864	LongSAGE
testis	NA	malignant pluripotent embryonal carcinoma (2102Ep cell line)	hs0213	Tag-seq
	NA	Embryonal carcinoma (32m; bulk)	1863	LongSAGE
	NA	malignant pluripotent carcinoma (22m; NTERA-2 cell line)	hs0212	Tag-seq

Group	Stage	Description	Library	Protocol
vascular	P0	Endothelial cells (3 male)	hs0237	Tag-seq
	P0	Endothelial cells (3 female)	hs0230	Tag-seq
	NA	Normal liver associated	655	LongSAGE
brain	NA	medulloblastoma (f cell line)	2083	LongSAGE
	NA	medulloblastoma (m cell line)	2085	LongSAGE
	NA	glioblastoma (stem cell line)	1644	LongSAGE
	NA	medulloblastoma (f cell line)	2123	LongSAGE
	NA	medulloblastoma (m cell line)	2129	LongSAGE
	NA	normal (m+f aborted fetuses)	656	LongSAGE
	NA	Normal substantia nigra	648	LongSAGE
	NA	glioblastoma (4d cell line)	1645	LongSAGE
	NA	glioblastoma (28d cell line)	1643	LongSAGE
breast	NA	Carcinoma (34f; extensive LCIS)	703	LongSAGE
	NA	Carcinoma epithelium (pleural effusion; recurrence)	2171	LongSAGE
	NA	Carcinoma epithelium (pleural effusion; recurrence)	2173	LongSAGE
	NA	Carcinoma epithelium (55f; invasive ductal)	2163	LongSAGE
	NA	Carcinoma epithelium (50f; invasive ductal)	2166	LongSAGE
	NA	Fibroadenoma (11f; benign neoplasia)	723	LongSAGE
	NA	Carcinoma associated stroma (28f)	650	LongSAGE
	NA	Breast carcinoma - white blood cells (47f; IDC7)	659	LongSAGE
	NA	Carcinoma epithelium (50f; invasive ductal)	2165	LongSAGE
	NA	Carcinoma epithelium (32f; invasive ductal)	2175	LongSAGE
	NA	Carcinoma epithelium (ascites; mixed lobular/ductal)	2169	LongSAGE
	NA	Carcinoma associated stroma (47f; IDC7)	646	LongSAGE
	NA	Carcinoma epithelium (47f; gradeII; IDC7)	645	LongSAGE
	NA	Phyllodes tumor fibroblasts (52f; malignant high grade; poorly differentiated)	683	LongSAGE
	NA	Carcinoma epithelium (tumor)	673	LongSAGE
	NA	Carcinoma (71f; bulk ductal invasive)	649	LongSAGE
	NA	Carcinoma epithelium (invasive ductal)	675	LongSAGE
	NA	Carcinoma (44f; bulk in situ ductal)	657	LongSAGE
	NA	Carcinoma associated myofibroblast (f; invasive ductal)	676	LongSAGE
	NA	Carcinoma associated myofibroblast (f; tumor)	674	LongSAGE
	NA	Carcinoma associated myofibroblast (47f; grade II; IDC7)	644	LongSAGE
	NA	Normal epithelium (22f; breast reduction; mammary epithelial stem cells)	2179	LongSAGE
	NA	Normal myoepithelium (47f; IDC7)	647	LongSAGE
	NA	Normal epithelium (22f; breast reduction; diff luminal epithelial cells)	2177	LongSAGE
	NA	Normal Stroma (44f; bulk)	1943	LongSAGE
	NA	Normal epithelium (22f; breast reduction)	2181	LongSAGE
colon	NA	Carcinoma (p53 KO; Anoxia)	654	LongSAGE
	NA	Carcinoma (p53 WT; Oxygen)	653	LongSAGE
	NA	Carcinoma (p53 WT; Anoxia)	651	LongSAGE
	NA	Carcinoma (p53 KO; Oxygen)	652	LongSAGE
esophagus	NA	Adenocarcinoma (cancer)	2147	LongSAGE
	NA	Dysplasia (low grade; pre-cancer)	2143	LongSAGE
	NA	Normal	2103	LongSAGE
	NA	Dysplasia (high grade; pre-cancer)	2145	LongSAGE

Group	Stage	Description	Library	Protocol
gall bladder	NA	Adenocarcinoma	2133	LongSAGE
	NA	Adenocarcinoma (tubular)	2131	LongSAGE
	NA	Normal (ventral wall)	2127	LongSAGE
	NA	Adenocarcinoma (tubular; poorly diff)	2125	LongSAGE
lung	NA	Adenocarcinoma (m; poorly diff)	963	LongSAGE
muscle	NA	Rhabdomyosarcoma	1923	LongSAGE
pancreas	NA	Normal	643	LongSAGE
retina	NA	Normal (central retina)	1993	LongSAGE
	NA	Retinoblastoma (bilateral; poorly diff; left orbit)	1883	LongSAGE
white blood cells	NA	Monocyte normal (39m; AP_A2)	1567	LongSAGE
	NA	Monocyte normal (71m; AP_P1)	1565	LongSAGE
	NA	Plaque macrophage normal (71m; AP_P1)	1983	LongSAGE
	NA	Monocyte normal (68f; AP_C1)	1564	LongSAGE
	NA	Monocyte normal (72f; AP_P2)	1563	LongSAGE
	NA	Lung macrophage normal (53m)	1987	LongSAGE
	NA	Monocyte depleted mononuclear cells normal (45m; AP_A1)	1569	LongSAGE
	NA	Monocyte depleted mononuclear cells normal (71m; AP_P1)	1568	LongSAGE
	NA	Plaque macrophage normal (72f; AP_P2)	1985	LongSAGE
	NA	Monocyte normal (45m; AP_A1)	1566	LongSAGE
	NA	Breast carcinoma (47f; IDC7)	659	LongSAGE

Appendix B CGAP library subgroups

For full table, see (ftp://ftp.bcgsc.ca/supplementary/ASMorrissy/Table_S4.pdf). Within each tissue, various experimental sub-groupings were made (E1-E9) that represented pairings of subsets of libraries into either A (cancerous), or B (normal). A and B are also used to distinguish libraries belonging to different cancer subtypes. Tag pair ratios were assessed between the libraries in A versus B categories, in each experimental group, for every tissue.

Group	Stage	Description	Experimental Sub-groups								
			E1	E2	E3	E4	E5	E6	E7	E8	E9
bladder	TaG2	Tumor (f)	A1		A3		A5				
	T4aG3	Tumor (m)	A1			A4					
	T1G3	Tumor (f)	A1	A2			B5				
	T1G3	Tumor (f)	A1	A2			B5				
	TaG2	Tumor (m)	A1		A3		A5				
	NA	Normal urothelium	B1	B2	B3	B4					
skin	Stage4	Metastatic melanoma (m)	A1	A2	A3				B7		
	NA	Squamous cell carcinoma (f)	A1	A2				A6			B9
	NA	Seborrheic Keratosis (f)	A1	B2		A4					
	NA	Seborrheic Keratosis (m)	A1	B2		A4					
	NA	Basal cell carcinoma (m)	A1	A2			A5				A9
	NA	Basal cell carcinoma (f)	A1	A2			A5				A9
	NA	Squamous cell carcinoma (m)	A1	A2				A6			B9
	NA	Normal nevus (f)	A1	B2		A4					
	Stage4	Metastatic melanoma (f)	A1	A2	A3				B7		
	NA	Normal nevus (m)	A1	B2		A4					
	Stage2	Primary melanoma (m; nonpigmented)	A1	A2	A3				A7		
	Stage2	Primary melanoma (m; pigmented)	A1	A2	A3				A7		
	NA	Normal skin (f)	B1	B2	B3	B4	B5	B6		A8	
	9months	Foreskin (m)	B1	B2	B3	B4	B5	B6		B8	
	NA	Normal skin (m)	B1	B2	B3	B4	B5	B6		A8	
uterus	NA	Ovary	B1								
	NA	Fallopian tube	A1								
bone marrow	NA	Peripheral blood from AML patient (m)	A1	A2							
	NA	Peripheral blood from AML patient (m)	A1	B2							
	NA	AML	B1		A3						
	NA	AML	B1		B3						

Appendix C Sense-antisense expression ratios

For full table, see (ftp://ftp.bcgsc.ca/supplementary/ASMorrissy/Table_S6.pdf). Normalized counts are shown for tags in each experimental subgroup. Tag counts are normalized to tags per million. In each tissue, two or more libraries were categorized as either M or N (and these were used to distingiush between normal and cancerous stages, or also between two different cancer stages). The number of M and N libraries in each group are shown, and tag counts are enumerated first in the M libraries, and then in the N libraries. Thus, for a group with 4 M and 2 N libraries, the first four counts represent those of the M libraries, and the last 2 counts represent those of the N libraries. Tags could belong to three types of genes: those with different isoforms (ISO), to S-AS genes (S-AS), or to Single-AS genes (Single-AS).

Gene id	Tag sequence	Tissue	Subgroup	M	N	Library descriptions and counts					
						TaG2	T1G3	T1G3	T4aG3	TaG2	Normal
ENSG00000151914	GAATCAAAGAGAAAGAT	bladder	E1	5	1	0	8.92	0	0	0	11.8
ENSG00000151914	TGAGGTTTTCTTTTGCT	bladder	E1	5	1	6.34	11.27	8.66	8.83	8.85	0
ENSG00000114861	CTTAGTCTAAAGACTGT	bladder	E1	5	1	11.48	6.26	6.5	6.73	7.9	0
ENSG00000114861	GTATGCAGAAATGTGAT	bladder	E1	5	1	7.32	0	0	0	6.01	11.68
ENSG00000146416	GTGGCGTGTGCCTGTAG	bladder	E1	5	1	20.22	0	3.61	0	14.39	24.4
ENSG00000146416	TGAAACTTTTCCTAGAT	bladder	E1	5	1	74.22	46.95	48.12	21.09	77.01	9.85
ENSG00000152465	AGGTCAGGAGATCGAGA	bladder	E1	5	1	370.86	36.31	48.72	60.06	38.21	159.81
ENSG00000152465	TCATACAGTTTGTGAAA	bladder	E1	5	1	4.15	0	39.22	10.49	9.08	0
ENSG00000189223	ACTCAATAAACCATTGC	bladder	E1	5	1	9.51	0	27.07	0	148.01	176.76
ENSG00000189223	ATGGCACCATATTGTGT	bladder	E1	5	1	7.76	0	17.2	0	0	0
ENSG00000210082	ACACAGCAAGACGAGAA	bladder	E1	5	1	0	0	0	198.95	156.03	70.68
ENSG00000210082	TGTCACTGGGCAGGCGG	bladder	E1	5	1	3.39	3.44	11.67	4.75	4.72	0
ENSG00000210082	ACACAGCAAGACGAGAA	bladder	E1	5	1	0	0	0	198.95	156.03	70.68
ENSG00000210082	CCTGTGTTGGGTTGACA	bladder	E1	5	1	8.96	3.91	32.48	13.69	11.44	0

Appendix D miRNA targeting sites

For full table, see (ftp://ftp.bcgsc.ca/supplementary/ASMorrissy/Table_S7.pdf). The frequency of unique miRNA targetting sites predicted by TargetScanS was assessed in all genes (All), in genes with differentially expressed isoforms between cancer and normal libraries (DE), and in the genes with isoforms with the 10% most extreme ratio changes (Top 10%). miRNA targetting sites with significantly enriched frequencies between the DE and Top 10% lists are shown (Benjamini and Hochberg corrected p-values; threshold of significance = 0.05).

miR-target	Top 10%	DE	All	p-value
miR-124.2/506	102	208	1021	1.86E-05
miR-181	78	154	631	6.47E-05
miR-200bc/429	71	139	536	9.06E-05
miR-224	38	62	218	9.06E-05
miR-15/16/195/424/497	60	114	628	1.12E-04
miR-128	61	118	610	1.68E-04
miR-19	65	129	629	2.15E-04
miR-377	39	69	260	5.27E-04
miR-495	62	125	543	5.27E-04
miR-203.1	50	96	422	6.01E-04
miR-17-5p/20/93.mr/106/519.d	58	117	645	7.60E-04
miR-93.hd/291-3p/294/295/302/372/373/520	40	73	411	7.60E-04
miR-141/200a	48	93	402	8.59E-04
miR-205	32	55	255	8.59E-04
miR-330	51	101	395	9.90E-04
miR-27	65	137	714	1.11E-03

Appendix E Functional annotation of known and novel SAS genes expressed in LCLs.

Functional annotations for (A) known SAS, (B) novel SAS genes, and (C) non SAS genes shows significantly enriched Gene Ontology terms and UniProt Keywords.

A

Category	Term	Gene Count	Fold Enrichment	p-value
4,792 Known SAS genes				
GO Biological Process	cellular protein metabolic process	601	12.9	1.70E-04
	regulation of small GTPase mediated signal transduction	87	1.9	2.10E-03
	protein modification process	379	8.1	3.30E-03
	RNA metabolic process	254	5.4	8.60E-03
	intracellular signaling cascade	328	7	9.50E-03
GO Cellular Component	cytoplasm	1797	38.5	4.40E-12
	cytoplasmic part	1198	25.7	2.80E-05
	intracellular membrane-bounded organelle	1874	40.1	7.60E-05
	intracellular organelle	2086	44.7	9.10E-05
	Golgi apparatus	249	5.3	2.30E-04
	DNA-directed RNA polymerase II, holoenzyme	36	0.8	1.50E-03
GO Molecular Function	Ras guanyl-nucleotide exchange factor activity	38	0.8	7.20E-03
	protein kinase activity	174	3.7	7.60E-03
	adenyl ribonucleotide binding	386	8.3	6.70E-03
UniProt Keywo`rds	alternative splicing	2037	43.6	3.70E-44
	phosphoprotein	1846	39.5	6.60E-19
	cytoplasm	875	18.7	3.70E-09
	coiled coil	551	11.8	1.20E-07
	guanine-nucleotide releasing factor	51	1.1	1.70E-04
	tpr repeat	60	1.3	1.30E-03
	golgi apparatus	172	3.7	1.90E-03
	atp-binding	348	7.5	5.40E-03

B

Category	Term	Gene Count	Fold Enrichment	p-value
7,648 Novel SAS genes				
GO Biological Process	negative regulation of gene expression	271	3.5	5.70E-06
	negative regulation of cellular metabolic process	370	4.8	3.90E-06
	negative regulation of transcription	248	3.2	5.60E-06
	negative regulation of macromolecule metabolic process	373	4.9	8.90E-06
	negative regulation of biosynthetic process	299	3.9	9.20E-06
	negative regulation of macromolecule biosynthetic process	287	3.8	7.80E-06
	negative regulation of cellular biosynthetic process	293	3.8	8.30E-06
	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	269	3.5	1.50E-05
	negative regulation of nitrogen compound metabolic process	271	3.5	2.40E-05
	regulation of protein metabolic process	283	3.7	2.60E-05
	cellular protein metabolic process	1070	14	4.00E-05
	negative regulation of RNA metabolic process	196	2.6	5.40E-05
	negative regulation of transcription, DNA-dependent	193	2.5	5.50E-05
	regulation of cellular protein metabolic process	245	3.2	2.00E-04
	neurogenesis	302	3.9	2.50E-04
	regulation of signal transduction	425	5.6	2.80E-04
	positive regulation of cell differentiation	129	1.7	4.00E-04
	generation of neurons	281	3.7	5.00E-04
	regulation of cell development	117	1.5	4.80E-04
	regulation of neurogenesis	95	1.2	3.70E-03
	regulation of cellular carbohydrate metabolic process	29	0.4	6.10E-03
	negative regulation of programmed cell death	184	2.4	6.30E-03
	embryonic limb morphogenesis	55	0.7	6.50E-03
	embryonic appendage morphogenesis	55	0.7	6.50E-03
	regulation of programmed cell death	385	5	6.60E-03
	limb morphogenesis	61	0.8	6.50E-03
	negative regulation of cell death	184	2.4	6.50E-03
	regulation of apoptosis	381	5	6.80E-03
	posttranscriptional regulation of gene expression	115	1.5	6.60E-03
	positive regulation of cellular carbohydrate metabolic process	18	0.2	7.20E-03
	positive regulation of carbohydrate metabolic process	18	0.2	7.20E-03
	positive regulation of RNA metabolic process	238	3.1	7.10E-03
	positive regulation of transcription, DNA-dependent	236	3.1	7.40E-03
	regulation of carbohydrate metabolic process	29	0.4	8.70E-03
	negative regulation of apoptosis	180	2.4	8.90E-03
	regulation of neuron differentiation	77	1	8.70E-03
	organ morphogenesis	274	3.6	8.50E-03
	positive regulation of cellular metabolic process	412	5.4	9.30E-03
	protein modification process	658	8.6	9.60E-03

Category	Term	Gene Count	Fold Enrichment	p-value
7,648 Novel SAS genes				
GO Cellular Component	cytoplasm	3273	42.8	3.50E-29
	cytoplasmic part	2226	29.1	1.30E-19
	intracellular organelle	3860	50.5	3.90E-19
	intracellular membrane-bounded organelle	3465	45.3	9.90E-19
	cytosol	692	9	1.10E-18
	intracellular organelle part	1885	24.6	5.00E-11
	nuclear part	845	11	3.30E-07
	nucleus	2196	28.7	4.10E-07
	intracellular organelle lumen	811	10.6	2.20E-05
	nuclear lumen	665	8.7	1.10E-04
	organelle envelope	305	4	1.40E-04
	nuclear envelope	115	1.5	2.50E-04
	Golgi apparatus	411	5.4	5.20E-04
	nucleoplasm	415	5.4	5.40E-04
	intracellular non-membrane-bounded organelle	1131	14.8	1.70E-03
	Golgi apparatus part	152	2	2.10E-03
	nuclear membrane	46	0.6	6.40E-03
	nuclear body	91	1.2	9.90E-03
GO Molecular Function	adenyl ribonucleotide binding	707	9.2	7.50E-08

Category	Term	Gene Count	Fold Enrichment	p-value
7,648 Novel SAS genes				
UniProt Keywords	phosphoprotein	3445	45	5.40E-94
	acetylation	1339	17.5	1.80E-44
	alternative splicing	3319	43.4	1.20E-44
	cytoplasm	1512	19.8	8.60E-20
	nucleotide-binding	806	10.5	7.00E-16
	atp-binding	641	8.4	3.50E-13
	nucleus	1830	23.9	7.70E-11
	metal-binding	1296	16.9	1.10E-09
	chromosomal rearrangement	158	2.1	4.00E-08
	ubl conjugation	297	3.9	4.40E-08
	activator	266	3.5	7.20E-08
	kinase	336	4.4	3.40E-07
	Transcription	901	11.8	3.50E-06
	transcription regulation	880	11.5	6.60E-06
	zinc	944	12.3	9.50E-06
	zinc-finger	752	9.8	1.50E-05
	repressor	217	2.8	2.20E-05
	rna-binding	262	3.4	2.50E-05
	host-virus interaction	150	2	2.90E-05
	protein biosynthesis	105	1.4	4.80E-05
	neurogenesis	85	1.1	5.80E-05
	isopeptide bond	163	2.1	8.70E-05
	cytoskeleton	299	3.9	9.80E-05
	transferase	611	8	1.20E-04
	sh3 domain	113	1.5	1.60E-04
	transport	720	9.4	2.20E-04
	Proto-oncogene	121	1.6	2.80E-04
	disease mutation	686	9	3.30E-04
	ATP	123	1.6	3.90E-04
	magnesium	211	2.8	7.70E-04
	ligase	152	2	7.80E-04
	coiled coil	853	11.2	7.50E-04
	golgi apparatus	272	3.6	8.80E-04
	endoplasmic reticulum	324	4.2	8.70E-04
	developmental protein	351	4.6	9.10E-04
	Endocytosis	57	0.7	1.10E-03
	calmodulin-binding	68	0.9	1.40E-03
	ubl conjugation pathway	236	3.1	2.20E-03
	mrna processing	129	1.7	3.30E-03
	Apoptosis	180	2.4	4.60E-03
	ribosome	44	0.6	5.40E-03
	ribonucleoprotein	136	1.8	5.50E-03
	phosphotransferase	102	1.3	5.70E-03
	mrna splicing	105	1.4	7.30E-03
	tyrosine-protein kinase	61	0.8	8.90E-03

C

Category	Term	Gene Count	Fold Enrichment	p-value
7,137 Non SAS genes				
GO Biological Process	defense response to bacterium	75	1.1	2.70E-11
	chemotaxis	82	1.1	1.30E-04
GO Molecular Function	chemokine activity	35	0.5	1.10E-06
	chemokine receptor binding	36	0.5	1.50E-06
	serine-type endopeptidase inhibitor activity	52	0.7	1.30E-04
UniProt Keywords	Secreted	809	11.3	1.30E-43
	signal	1359	19	1.80E-36
	Antimicrobial	57	0.8	1.70E-13
	antibiotic	55	0.8	2.30E-13
	cytokine	107	1.5	2.20E-11
	defensin	36	0.5	6.60E-11
	disulfide bond	1101	15.4	1.10E-10
	hormone	59	0.8	3.40E-09
	protease inhibitor	64	0.9	8.20E-07
	cleavage on pair of basic residues	132	1.8	1.00E-06
	chemotaxis	45	0.6	2.70E-05
	inflammatory response	46	0.6	1.10E-04
	Lectin	80	1.1	1.60E-04
	Serine protease inhibitor	46	0.6	3.90E-04
	Intermediate filament	45	0.6	4.20E-04
	immune response	104	1.5	4.00E-04
	plasma	50	0.7	1.40E-03
	neuropeptide	24	0.3	2.10E-03
	inflammation	19	0.3	2.20E-03
	Monooxygenase	43	0.6	2.30E-03
	amidation	28	0.4	2.40E-03
	fungicide	11	0.2	3.60E-03