12

Alternative expression analysis: Experimental and bioinformatic approaches for the analysis of transcript diversity

Malachi Griffith and **Marco A. Marra**[1]

Introduction

The human genome contains approximately 30,000 protein coding genes (Venter *et al*., 2001; Lander *et al*., 2001). These loci generate the functional components of the cell and represent ~1% of the complete genomic sequence. Although the human genome sequence itself provides a crucial framework for the study of biology, understanding the function of genes requires analysis of the 'transcriptome' encoded by the genome and the 'proteome' it gives rise to. Gene transcription occurs in the nucleus followed by capping, RNA splicing, polyadenylation and export to the cytoplasm. Transcription thus involves three related processes which collectively define the ultimate sequence content of each transcript. First, an RNA polymerase binds to a transcriptionally competent 'unwound' region of genomic DNA template and results in the synthesis of a pre-mRNA molecule in the 5′ to–3′ direction. RNA polymerase II transcribes most human genes and initiates transcription at specific positions in the genome called transcription initiation sites which are found downstream of promoter elements recognized by transcription factors. The initiation site chosen by the polymerase defines the 5′ end of the resulting transcript. Second, RNA splicing results in the removal of most of the nucleotides of the pre-mRNA transcript. Splicing involves the recognition of splice sites, removal of introns from a pre-mRNA transcript and joining of adjacent exons to yield a mature mRNA transcript (Plate I). The splicing process is mediated by

[1] British Columbia Cancer Agency, Genome Sciences Centre, 675 West 10th Avenue, Vancouver, British Columbia, V5Z 1L3, Canada.

a series of protein-protein, RNA-protein, and RNA-RNA interactions involving a number of sequence motifs in addition to the actual splice sites (Black, 2003). The splice sites chosen during this process define the primary structure of the resulting transcript. Finally, the 3′ end of the transcript is defined by polyA polymerase which cleaves the transcript and adds a poly-A tail 10 to 30 nucleotides downstream from a recognition site in the RNA transcript. For years, these three processes were thought to occur in a prescribed way for each gene and deviations from the 'one-gene-one-product' model were considered rare.

A major challenge in decoding the information content of the human genome is presented by the processes of alternative transcription (AT), which can produce from a single locus, transcripts with different combinations of exons. More precisely, alternate transcripts may arise from a single locus by the use of alternative transcription initiation (ATI), alternative splicing (AS) and alternative polyadenylation (AP) sites. The mechanisms by which these sites are selected by the transcription machinery are tightly coupled to each other (Kornblihtt, 2005; Matlin *et al*., 2005; Maniatis and Reed, 2002), involving many of the same protein and RNA factors and will be broadly examined as different facets of the same biological phenomenon throughout this chapter. The idea that alternative transcription dramatically increases the functional diversity of the proteome has gained general acceptance in recent years (Black, 2000; Lareau *et al*., 2004; Maniatis and Tasic, 2002; Roberts and Smith, 2002). Based on an analysis of ~1.4 million sequenced human clones it is estimated that approximately 52% of human genes utilize alternate transcription initiation sites (Suzuki *et al*., 2004). Similarly, recent estimates suggest that as many as 74% of human genes undergo alternative splicing, a process which can produce multiple transcripts with different combinations of exons from a single gene locus (Johnson *et al*., 2003). Alternative splicing produces distinct isoforms by a number of modes including: exon skipping, use of alternate mutually exclusive exons, use of alternate 5′ or 3′ splice sites and the retention of intronic sequences (Plate I and II; Kalnina *et al*., 2005). Recognition of a particular exon by the splicing machinery is mediated by splicing acceptor and donor sites which define the boundaries of each exon as well as by exonic and intronic splicing enhancers and silencers (Black, 2003; Berget, 1995). Finally, a recent annotation of the transcripts for ~8,000 human genes in the 'AltTrans' database suggests that ~60% of human genes utilize alternate polyadenylation sites (Le Texier *et al*., 2006). Plate I and II summarize the types of alternative transcription sites and some of the surrounding motifs which influence their selection by the transcriptional machinery. A current challenge of genome research is to catalogue all possible transcriptional outcomes for every gene; to define the pattern of expression of these transcripts associated with development, tissue and disease states; and to determine the regulatory networks which control these patterns. A detailed description of the regulation of these processes is beyond the scope of this chapter, but excellent reviews on the mechanisms of regulation and the experimental methods used to study them are available (Black, 2003; Soller, 2006; Cooper, 2005; Hicks *et al*., 2005; Ule *et al*., 2005).

Based on the apparent prevalence of alternate transcript initiation sites, splice sites, and polyadenylation sites, the number of proteins encoded by the human genome

is likely to be much greater than the number of gene loci and has been estimated to be as high as 100,000 (Goldstrohm *et al*., 2001; Harrison *et al*., 2002). The biological consequences of this observation are significant. AT appears to be an important mechanism for encoding a diversity of functions at a single genomic locus and this diversity may be realized in part through alterations in protein-protein interactions and subcellular localization. Mutations or polymorphisms in the genes responsible for transcription initiation, splicing and polyadenylation may affect the transcriptional outcome of many genes and contribute to disease (a 'trans-acting' effect) (Srebrow and Kornblihtt, 2006). Similarly, inherited or acquired mutations and common polymorphisms within the sequence motifs which regulate these processes for each individual gene could also contribute to disease (a 'cis-acting' effect). Thus, to effectively characterize the human transcriptome and apply this knowledge to problems of medical significance, it is necessary to document the prevalence of AT and consider the biological roles of proteins encoded by alternative transcripts.

Until recently it was not possible to measure the prevalence of AT or detect comprehensively the diverse transcripts produced by it. With the availability of high-density microarrays and the advent of next-generation sequencing technologies, there is now an opportunity to study AT on a genome-wide scale. The implications of these technical developments and their application to the study of AT are substantial, for up until this time measurements of gene expression relied largely on the detection of a single transcript for each gene. Microarrays designed to detect differential AT will drive the discovery of transcripts with novel, functionally relevant exon combinations, and such discoveries will inform on the protein coding potential of metazoan genomes. Similarly, ready access to sequence data for multiple transcripts from a single locus will provide invaluable validation of their precise sequence content. In addition to fueling basic research questions, it is easy to imagine how knowledge of the transcripts and proteins produced by AT could lead to medically relevant discoveries. For example, novel exon combinations expressed in disease states might yield excellent candidates for development of new diagnostic tools and therapies.

Having introduced what is meant by the term 'alternative transcription' and described how a single locus can produce multiple distinct transcripts, the remainder of this chapter will address the following areas: (1) the experimental and bioinformatic approaches currently available to comprehensively profile transcript diversity and what these methods have revealed about the prevalence and nature of AT, (2) the functional significance of AT and (3) the implications of AT for the study of disease.

Genomic approaches for the study of transcript diversity

The prevalence and perceived importance of AT has increased dramatically over the last two decades. For example, early estimates suggested that alternative splicing was a relatively unusual event occurring in approximately 5% of all genes (Sharp, 1994). The advent of genome-wide studies of transcript diversity, involving the analysis of short expressed sequence tags (ESTs) by alignment to the genome and annotation of the exons present have resulted in predictions that at least 42% of human genes exhibit

AS (Huang *et al*., 2003). Such studies have also resulted in the creation of a number of databases of observed initiation, splicing, and polyadenylation events as well as the identification of AT regulatory motifs for a number of species (Table 1). More recently, exon-junction microarray experiments used to survey splicing events in 52 human tissues and cell lines found that as many as 74% of all human genes are alternatively spliced (Johnson *et al*., 2003). The rationale for conducting such efforts is that the determination of gene function and identification of therapeutic targets can be improved by first determining the subset of genes and isoforms which are actually expressed in relevant tissues and disease states. Preliminary experiments suggest that AT occurs most frequently in tissues with diverse cell types such as brain, metabolically active tissues such as testis and liver and cell types with highly diversified functions such as immune cells (Modrek *et al*., 2001; Yeo *et al*., 2004; Watson *et al*., 2005; Noh *et al*., 2006). The following sections will describe the computational and experimental ways in which transcript diversity can be studied by using genomic DNA sequence, cDNA library sequencing, tag-based library sequencing, microarray approaches, and finally methods for the visualization and functional validation of alternative transcripts. The advantages and disadvantages of each of these approaches are summarized in Table 1. Each method is depicted in Plate III and IV.

In silico methods

One starting point for the analysis of a species' transcriptional units (which generally correspond to genes) and often one of the first large sources of data for that species is the genome sequence itself. Perhaps the most important issue faced in analyzing the transcript diversity generated by a particular genome is the problem of accurate and reliable annotation of the genes present. Several algorithms which attempt to annotate the genome by predicting gene structure have been described (Jones, 2005). Generally these predict a single transcript per gene but some have been adapted to consider the occurrence of multiple alternative transcripts generated from a single locus. A few computational methods have also been recently developed specifically to predict AT directly from genomic sequences without the use of experimentally derived expression data. For example, methods have been developed for the prediction of exon skipping events by considering only the genomic sequence of an exon in the human genome and its ortholog in another species such as mouse (Sorek *et al*., 2004; Yeo *et al*., 2005; Flicek and Brent, 2006). This approach is aided by the fact that the sequence of alternative exons and the flanking intronic sequence exhibit generally higher levels of conservation between related species than the sequence of 'constitutive' exons (those found in every transcript) (Modrek and Lee, 2003; Sorek and Ast, 2003). Each of these methods generally requires a training set of a few thousand known exon skipping events that are conserved between human and mouse. Although these methods are capable of predicting exon skipping events based solely on the genomic sequence of human and mouse, the data sets used to train them are derived from previously observed expressed sequence tags (ESTs). The training set is used to develop a model by which a 'signature' or classifier is generated to enable prediction of skipped exons across the

Table 1. Summary of methods for studying transcript diversity

Method	*Events detected*	*Description (strengths/limitations)*
Computational methods		
Predict transcription events from genomic sequence without using expression data.		
Ab initio	ATI, AS, AP	Predictions based on a single reference genome, not quantitative, low sensitivity/specificity compared to methods that use expression data.
Comparative genomic	ATI, AS, AP	Predictions rely on existence of suitable comparative genomes, not quantitative, medium sensitivity/specificity compared to methods that use expression data.
Sequence-based methods		
Generate expressed sequence data from RNA, align to genome and annotate transcription events. These methods do not rely on pre-existing gene annotations and they are capable of providing exon boundary/connectivity information as well as novel gene discovery.		
EST cDNA	ATI†, AS†, AP	End bias, partial transcripts (300–1000 bp reads), high cost, medium throughput, limited quantitative value.
FL-cDNA	ATI, AS, AP	Complete transcripts, high cost, low throughput, results in a physical copy of transcript, not quantitative.
Targeted FL-cDNA	AS	Near complete transcripts, high cost, low throughput, results in a physical copy of transcript, not quantitative.
SAGE	AS†, AP	3′ end bias, short tags (17–21 bp), medium cost, medium throughput, quantitative.
CAGE	ATI	5′ end bias, short tags (20 bp), medium cost, medium throughput, quantitative.
GIS	ATI, AP	End bias, short tags (40 bp paired end tags), medium cost, medium throughput, quantitative.
SOLEXA SBS	ATI, AS, AP	Short tags (16–20 bp), low cost, high throughput, quantitative.
454/Roche GS20 SBS	ATI, AS, AP	Short tags (~100 bp), low cost, high throughput, quantitative.
Microarray-based methods		
Fluorescently label RNA and hybridize to an array of 'spots' each representing content from a reference genome. All array methods are subject to cross-hybridization between related sequences.		
Spotted cDNA	None	Limited to composition of cDNA library, not capable of distinguishing transcript variants, low cost, high throughput, quantitative.
3′ Expression	AS†, AP†	3′ end bias, limited by pre-existing gene annotations, low cost, high throughput, quantitative.
Whole genome tiling	ATI†, AS†, AP†	Not limited by pre-existing gene annotations, potential for gene discovery, high cost, quantitative.
Exon tiling	ATI†, AS†, AP†	Limited by pre-existing gene annotations, low cost, high throughput, quantitative.
Splicing arrays	ATI, AS, AP	Limited by pre-existing gene annotations, provides exon boundary/ connectivity information, medium cost, medium throughput, quantitative.

Abbreviations: (ATI) alternative transcript initiation; (AS) alternative splicing; (AP) alternative polyadenylation; (SBS) sequence by synthesis; (†) limited applicability or supporting evidence.

entire genome. Experimental validations of the predictions of these methods have revealed a sensitivity value as high as 73% at 64% specificity (Sorek *et al*., 2004). Based on the simple assumption that alternatively transcribed exons will be highly conserved and surrounded by highly conserved intronic sequences, it is also possible to accurately predict such events based solely on the genomic sequence of related species without use of an EST training set. Philipps *et al. (*2004) used this approach to identify alternative exons representing all of the major classes of AS in *Drosophila* by comparing the genomic sequence of *D. melanogaster* and *D. pseudoobscura*. The authors were able to confirm AS in 25% of the predicted alternatively spliced exons generated from this approach by RT-PCR whereas only 3% of randomly selected exons were found to be alternatively spliced. The pool of alternative exons that were confirmed in this experiment was found to be enriched for exons that preserve the reading frame of the predicted protein and the highly conserved intronic sequence surrounding these exons was found to be larger than in constitutive exons. Since these initial reports, more sophisticated methods for distinguishing alternative exons from constitutive exons have emerged. For example, a support vector machine (SVM) learning procedure was used to develop a classifier for identification of alternative exons based on seven major exon attributes (exon size, divisibility by 3, conservation, splice site strength, etc.) and several additional minor attributes (Dror *et al*., 2005). This approach achieved a sensitivity of 50% with a corresponding specificity of 99.5% for human exons. Methods that are conceptually similar to this approach but use a hidden Markov model (HMM) instead of an SVM to identify alternative exons have also been described (Cawley and Pachter, 2003; Ohler *et al*., 2005). One of the problems faced by all conservation based AT prediction approaches is that they are difficult to implement for small exons and they are incapable of predicting species-specific events.

A recently developed algorithm, 'AUGUSTUS', has been proposed as the first purely *ab initio* method for gene prediction. This method is capable of predicting multiple transcripts for a gene from the sequence features of a single underlying genomic sequence without using conservation between sequences or expression data (Stanke *et al*., 2006). Xia *et al*. (2006) also recently described a purely *ab initio* method for identifying alternative splice sites which uses a model of predicted competition between neighboring splice sites to classify exons as either constitutive or alternative based on their genomic sequence alone. Although these approaches may be useful for analysis of species where very little expression data or suitable comparative genomes are available, in general such methods perform poorly compared to those that can incorporate comparative genomics and expression data.

Library construction and sequencing methods

ESTs sequencing of cDNA libraries

The earliest large repositories of data on transcript diversity consisted of expressed sequence tags (ESTs) generated by single sequence reads from systematically selected cDNA clones. Construction of a cDNA library commonly involves extraction of total RNA from cells, purification of polyA+ mRNAs, RT-PCR with an oligo d(T) primer

and cloning into a convenient vector. The rapid generation and sequencing of these libraries from specific human tissues became common in the early 1990's and rapidly accelerated the discovery and annotation of novel genes (Adams *et al*., 1993; Hillier *et al*., 1996). EST libraries are generally derived from a single normal or diseased tissue sample or a small pool of tissue samples. Most EST records deposited in public databases contain information on the tissue source and disease status of the sample from which they were derived. Typically each EST represents either the 5′ or 3′ end of a clone and initially the lengths of these reads were 300–500 nucleotides (Plate IV). Although improvements in Sanger sequencing have approximately doubled this read length, the majority of all ESTs do not represent a complete cDNA sequence and the overall coverage of EST data is heavly biased towards the 3′ end of transcripts (Takeda *et al*., 2006). The completion of the human genome, the comprehensive sequencing of EST libraries from a variety of tissues, and the continuing development of algorithms for 'spliced' alignments such as Blat, Spidey and Sim4 has allowed a first comprehensive assessment of the diversity of transcription (Table 2). EST sequences can be rapidly generated and aligned to a reference genome allowing the annotation of exon-intron boundaries and the inference of underlying transcript isoforms (Xie *et al*., 2002; Kim *et al*., 2005). Protein coding information may also be incorporated into predictions by performing 6-frame translations. The size of an EST library has been historically as small as a few hundred sequences or as large as tens of thousands and in rare cases even larger. Since a single cell type is likely to express 10 to 30k genes with a total of approximately 300 to 500k mRNA molecules per cell, the coverage of these EST libraries is not likely to provide an accurate quantitative measure for the expression of genes in a bulk tissue sample, especially given the fact that a majority of all transcripts will be derived from a minority of loci (Schmitt *et al*., 1999). The problem of over-representation of highly expressed genes can be addressed by applying normalization or subtraction techniques during the library construction phase (Bonaldo *et al*., 1996). These techniques enhance the rate of gene discovery but reduce the quantitative value of the data generated from such libraries. These approaches can also have the side effect of reducing the presence of transcript variants with subtle but potentially important variations and estimates of AT prevalence in the genome are likely to be underestimates as a result. To date, approximately 39 million EST sequences have been deposited in the public repository dbEST, and 7.9 million of these were generated from human samples (www.ncbi.nlm.nih.gov/dbEST/) (Boguski *et al*., 1993). This collection represents an incredible source of independent transcription observations from a wide variety of tissues and it has been used to identify differentially expressed genes specifically associated with particular tissues or disease states (Schmitt *et al*., 1999). Perhaps the most prominent examples of the use EST sequencing are the Cancer Genome Anatomy Project, which has attempted to create a complete catalogue of genes expressed in normal and cancerous tissues, and Unigene, which attempts to group all such sequences into clusters that appear to be expressed from a single locus (Strausberg, 2001).

Analysis of ESTs has proved to be a rich source for discovery of novel genes and transcript diversity and has led to a number of interesting observations about

Table 2. Alternative transcription resources

Resource name	*Description (applicable species)*	*Reference*
Ab initio/de novo alternative transcript prediction		
AUGUSTUS	Prediction of ATI, AS, and AP using only human genome sequence	Stanke *et al*., 2006
MARS	Human AS transcript prediction from pairwise alignments of mouse, rat, dog, opossum and frog genomes	Flicek and Brent, 2006
Spliced alignment algorithms (Churbanov et al., 2005)		
BLAT, SIM4, SPA, SPIDEY, Splice Predictor TAP	Identification of splice sites and gapped alignment of mRNAs to a reference genome	Kent, 2002; Florea *et al*., 1998; van Nimwegen *et al*., 2006; Wheelan *et al*., 2001; Usuka *et al*., 2000; Kan *et al*., 2001
Databases of transcript diversity derived from EST/mRNA sequences		
AltTrans	Annotation and visualization of AS and AP (Hs, Mm)	Le Texier *et al*., 2006
ASAP II	Annotation and visualization of AS (15 species)	Kim *et al*., 2006
ASD	Annotation and visualization of AS (Hs, Mm)	Stamm *et al*., 2006
ASPIC	Annotation and visualization of AS (Hs, Mm, and 15 other species)	Castrignano *et al*., 2006
ATID	Manual and computational annotation of ATI (Hs, Mm and 32 other species)	Cai *et al*., 2005
DBTSS	Database of ATI (Hs, Mm, zebrafish, etc.)	Suzuki *et al*., 2004
ECgene	Functional annotation of AS (Hs, Mm, Rn, etc.)	Kim *et al*., 2005
Hollywood	Annotation and visualization of AS (Hs, Mm)	Holste *et al*., 2006
LSAT	ATI, AS, and AP extracted from literature by text mining	Shah *et al*., 2005
MAASE	Manual annotation of AS (Hs, Mm)	Zheng *et al*., 2005
PolyA_DB	Annotation and visualization of AP (Hs, Mm)	Zhang *et al*., 2005
SpliceInfo	Annotation and visualization of AS (Hs)	Huang *et al*., 2005
TISA	Annotation of tissue specific transcripts (Hs, Mm)	Noh *et al*., 2006
T-STAG	Annotation of tissue specific transcripts (Hs, Mm)	Gupta *et al*., 2005
Alternative transcription regulatory element prediction (Zhang et al., 2005)		
ESEfinder	Identification of ESE sites and predicted effect of mutations within them	Cartegni *et al*., 2003
GRSDB	Identification of G-rich (GRS) processing motifs	Kostadinov *et al*., 2006
RegRNA	Identification of transcription and splicing regulatory sequences within RNAs	Huang *et al*., 2006
RESCUE-ESE	ESE annotation tool (Hs, Mm, zebrafish, pufferfish)	Fairbrother *et al*., 2004
TassDB	Collection of tandem splice sites (human, mouse, etc.)	Hiller *et al*., 2006
Validation/Visualization Tools		
ASePCR	Electronic PCR utility for validation of alternate isoforms	Kim *et al*., 2005
ASGS	Web based tool for AS graphs	Bollina *et al*., 2006
ASTRA	Visualization and classification of transcription patterns	Nagasaki *et al*., 2006
VISTA, UCSC, EnsEMBL	Generic browsers for visualization of expression data and comparative genomics	Frazer *et al*., 2004; Kuhn *et al*., 2006; Hubbard *et al*., 2005

Abbreviations: (ATI) alternative transcript initiation; (AS) alternative splicing; (AP) alternative polyadenylation; (ESE) exonic splicing enhancer; (Hs) *Homo sapiens*; (Mm) *Mus musculus*; (Rn) *Rattus norvegicus*.

transcription. Early analyses suggested that most AS events affect the 5′ UTR of genes, occur in at least 35% to 42% of all genes (Mironov *et al*., 1999; Modrek *et al*., 2001), and seem to be more prevalent in humans than in other species considered to date (Brett *et al*., 2002). Furthermore, within humans, the prevalence of AT varies dramatically between tissues. Brain and testis have the most exon-skipping events and liver has the most alternate splice site usage but one of the lowest rates of exon skipping (Yeo *et al*., 2004). Certain protein domains seem to be preferentially affected by AT and more than 50 domains that are commonly removed by AT have been identified (Resch *et al*., 2004). Analysis of these domains indicates that one of the central roles of AT may be to modulate protein-protein interactions. A number of groups have used ESTs to create databases of annotated AT events and characterize some of the general features of transcription diversity in metazoan species (Table 2). Among the results of these studies were the observations that skipped exons tend to be shorter than constitutively spliced exons, retained introns are generally shorter than those that are constitutively spliced, the introns flanking skipped exons tend to be longer, skipped exons are more likely than constitutively spliced exons to have a length that is a multiple of three, splice sites corresponding to constitutively spliced events tend to more closely resemble the consensus sequence than those involved in AS events, and the average sequence conservation between human and mouse is greater for alternatively spliced exons than constitutively spliced exons.

Full-length sequencing of cDNA libraries

As the cost of Sanger sequencing and primer synthesis has gone down it has become more practical to conduct full length sequencing of cDNA clones representing complete transcripts (Plate III). This is conceptually the simplest approach to study transcript diversity because it involves the capture and complete sequencing of single cDNAs. The complete structure of the transcript including the presence of alternative exons is thus determined. Large scale cDNA sequencing projects such as those associated with the Mammalian Gene Collection (MGC) and Full-length Long Japan (FLJ) projects are at various stages of completion for human, mouse and other species (Gerhard *et al*., 2004; Okazaki *et al*., 2002; Ota *et al*., 2004). The cDNA libraries for these efforts are generated in a similar way as that employed for EST sequencing but additional emphasis is placed on the generation of 'full-ORF' cDNAs. Sequencing of these cDNA clones involves generating EST end reads followed by sequencing of any remaining unknown portion by primer walking or transposon mediated sequencing (Butterfield *et al*., 2002). The resulting reads are then assembled into a contiguous sequence representing the entire mRNA. Initially, clones were selected for full-length sequencing by first generating EST end reads and identifying a subset of non-redundant clones. Although the primary goal of the MGC is to create a physical resource of cDNA clones for the analysis of gene function, the process of rescuing and sequencing these clones has led to considerable discovery of transcript diversity. More recently, the random clone sequencing approach of MGC has been replaced by an RT-PCR targeted approach in which amplicons for a known target gene are generated, cloned and sequenced (Baross

et al., 2004). The random clone sequencing approach has the potential to identify transcripts that differ in their transcription initiation, polyadenylation, and splicing. Because the targeted approach pre-defines the expected ends of the transcript it is only capable of detecting splice variation that occurs within these boundaries. However, since the cloning and rescue process generates many clones per target sequence, novel transcript variants of this type are routinely observed (Plate III). The MGC collection currently contains clones for ~15,000 genes generated from over one hundred tissue libraries. A recent study of ~56,000 full-length human clone sequences from the 'H-invitational human transcriptome' annotation meeting (Imanishi *et al*., 2004) found that these clones could be mapped to ~24,000 loci and 41% of these loci were represented by multiple cDNAs (Takeda *et al*., 2006). Of these loci, where at least a preliminary assessment of transcript diversity was possible, 68% showed evidence of AT with an average of approximately three unique transcripts per locus. Of these transcripts, 45% exhibited exon skipping events, 52% used at least one alternate 5′ or 3′ splice site, 15% had retained introns, and 3% used one of a series of mutually exclusive exons. Only 14% of the intron retention events were predicted to result in a transcript possibly subject to nonsense mediated decay (NMD). The majority (73%) of alternate transcripts exhibited a splicing event within the CDS of the predicted protein but 26% had events confined to the 5′ UTR and 6% had events confined to the 3′ UTR. Furthermore, if the rate of each type of event relative to the number of exons in each of these regions is calculated, events affecting the 5′ UTR have the highest frequency. Of all genes with observed AT events, 44% had events which occurred within a known protein motif, 44% were predicted to affect subcellular localization, and 20% were predicted to affect a transmembrane domain. Although the majority of human gene loci (59% in the study above) are still represented by only a single clone sequence, this initial data will act as a foundation for future studies of the diversity of transcripts generated from these loci. Similar analyses of almost 200,000 publicly available full length clone sequences derived from ~200 mouse tissues have resulted in similar findings to those observed in human. At least 40–70% of mouse genes have evidence for AT (Hayashizaki and Carninci, 2006; Okazaki *et al*., 2002; Takeda *et al*., 2006; Zavolan *et al*., 2003) and an estimated 78,000 distinct proteins are transcribed from only ~20,000 loci (Carninci *et al*., 2005). As described for the analysis of large EST datasets, these studies are invaluable for identifying the types of alternative transcripts that occur, revealing patterns in the size distribution, sequence composition and conservation of alternatively transcribed exons themselves and predicting their effect on resulting proteins (Zheng *et al*., 2005; Zavolan *et al*., 2003).

The complexity of the mammalian transcriptome generated by AT has been accepted as an outstanding challenge and was specifically discussed at the outset of the Mammalian Gene Collection project which has focused on the goal of acquiring a single 'representative' transcript for each known gene (Strausberg *et al*., 1999). Creating a comprehensive annotation of the complete mammalian transcriptome remains a daunting challenge and creating a physical collection of every transcript variant of every gene is currently unfeasible. Although methods that involve RT-PCR, cloning and sequencing of alternate transcripts can be accurate and revealing about the structural

differences of alternate isoforms, they are costly and may be difficult to scale up. Efforts to use EST and cDNA approaches to profile transcription in contrasting samples therefore have been limited to relatively small numbers of samples and have focused on gene annotation and transcript variant discovery rather than quantitative profiling of expression levels. The limited scope of these methods is insufficient to provide robust identification and quantification of alternatively spliced variants across samples representing a large number of tissues, patients or disease states. Bioinformatic analyses of all publicly available EST data are more comprehensive but are limited by the coverage of existing libraries and other problems such as end bias. The EST and cDNA libraries that are publicly available were not specifically intended to provide an accurate and consistent comparison of tissues or the progression of disease states, and often represent pools of individuals or cell types. Furthermore, although the use of EST and cDNA data to study splicing can be effective and has led to significant advances in our knowledge of AS it remains expensive and time consuming to create and sequence libraries of sufficient depth to quantitatively survey the transcripts present in samples representing several conditions.

A number of experimental approaches have recently been developed to specifically enrich libraries for alternative transcripts and thus increase the discovery of novel transcripts. One approach involves the construction of alternative splicing libraries (ASLs) representing differentially expressed exons from contrasting biological samples (Watahiki *et al*., 2004). Briefly, this protocol involves creating two cDNA libraries from cytoplasmic RNA, one from each of the samples to be compared. These two libraries are then processed such that single stranded sense DNA molecules are generated from one library and single stranded antisense DNA molecules are generated from the second library. The two libraries are then mixed to allow hybridization and formation of heteroduplex or 'loop' structures. This can occur in the event that a transcript from one library contains exon content not found in the corresponding transcript present in the second library. Hybrid molecules containing these loop structures are then selectively captured with biotinized random 25-mers which are purified on streptavidin conjugated magnetic beads and the resulting alternative transcript enriched cDNA population is cloned and sequenced. Use of this approach to compare melanocyte and melanoma cell lines identified 662 AS events representing all of the major categories of AS and differential splicing between the two cell lines was confirmed by RT-PCR for 73% of candidate exons. A comparison of this library construction approach to one without the splicing selection step suggested a ~40-fold enrichment for AS events. Thill *et al*. (2006) recently described a similar method, 'ASEtrap' for the construction of libraries enriched for alternative splicing events from a single RNA sample (rather than from a comparison of two samples). This method also utilizes the formation of loop structures in cDNA heteroduplexes caused by alternative transcripts of a single gene within the sample. These loops are captured by a recombinant *Escherichia coli* single-stranded DNA binding protein and then cloned and sequenced. Comparison of ~10,000 sequences generated from either an ASEtrap library or a control library revealed a ~10-fold enrichment for AS events in the ASEtrap library. A third approach for enrichment of AS isoforms (EASI) was recently proposed as a simpler version of the ASEtrap method

which can be rapidly employed to comprehensively profile all of the isoforms of a single target gene (Venables and Burn, 2006).

Sequence-tag based methods

The simplest way to overcome the issues of cost, poor representation of rare transcripts and lack of quantitative power in sequence based methods such as EST and full-length cDNA sequencing is to increase the number of sequences available for analysis. Serial analysis of gene expression (SAGE) (Velculescu *et al*., 1995; Saha *et al*., 2002) has been used as an alternative to EST sequencing and libraries as large as several hundred thousand tags have been produced (Boon *et al*., 2002; Siddiqui *et al*., 2005). Briefly, SAGE involves double stranded cDNA synthesis with an oligo(dT) primer, followed by digestion of the resulting cDNA with a restriction enzyme predicted to result in at least one cleavage per transcript (typically NlaIII). The resulting fragments are captured at the 3′ end by oligo(dT) primers coupled to streptavidin beads, and a type II restriction enzyme (e.g., MmeI) is used to create fragments of a fixed length (up to 21 bp) which are concatenated, cloned into a vector and sequenced. Each sequence read thus produces 30–45 tags corresponding to the 3′ end of transcripts from which they were derived (Plate III). By generating large numbers of these reads, a quantitative and digital form of expression data is produced with the number of tags mapped to each genomic locus representing the expression level of that gene. This form of data has been shown to have a moderate to low correlation (r = 0.5–0.8) of expression values when compared to microarray based approaches (Lu *et al*., 2004; van Ruissen *et al*., 2005). Two of the largest initiatives to make use of this technology are the Cancer Genome Anatomy Project (Boon *et al*., 2002) and the Mouse Atlas of Gene Expression Project (Siddiqui *et al*., 2005) each producing several million tags from a wide range of cell types for human and mouse respectively. Analysis of these large datasets has resulted in the identification of differentially expressed genes associated with disease, development or a specific tissue as well as the discovery of novel genes and transcript variants. An analysis of the SAGE tags mapping to ~13,000 EnsEMBL genes produced a prediction that 64% of genes exhibit AT and many of the variants observed were significantly differentially expressed in specific tissues or developmental stages in mouse (Siddiqui *et al*., 2005). Several bioinformatic tools to assist in the analysis and visualization of SAGE data have been recently developed (Boon *et al*., 2002). The tool, 'SAGE2Splice' was specifically designed to identify novel splice junctions in SAGE tags but is limited in its ability to profile exon connections by the fact that only 5–6% of tags span a splice site (Kuo *et al*., 2006). The disadvantages of SAGE include the theoretical occurrence of multiple tags per gene from incomplete digestion and the short length of each tag, both of which complicate the process of mapping tags to the gene from which they were expressed. Distinguishing tag artifacts created by mis-priming during library creation from tags derived from the use of alternative polyadenylation sites, alternative splicing or polymorphisms in restriction enzyme sites is also potentially problematic. Finally, because SAGE library construction involves the capture of tags corresponding to restriction enzymes sites closest to the 3′ end of each transcript, any variation observed is heavily biased towards the 3′ end of genes.

A complementary approach to SAGE, cap analysis of gene expression (CAGE), is used in a similar way as SAGE to profile the 5′end of transcripts and thereby acts as a means of identifying alternate promoter usage (Shiraki *et al*., 2003). Briefly, transcripts are captured by their 5′ cap (a modified guanosine nucleotide) and used to generate DNA tags of 20 nucleotides in length which are concatenated, cloned and sequenced. Each sequenced tag corresponds to the 5′ end of a single mRNA transcript and as with SAGE, the short length of each tag allows an increase in throughput and therefore depth of sampling and corresponding reduction in cost. By capturing many tags from a single gene the use of alternate transcription initiation (ATI) sites and their corresponding promoters can be catalogued. Generally 55–65% of sequenced tags can be unambiguously mapped to the genome (Shiraki *et al*., 2003). A recent analysis of 7.2 and 5.3 million CAGE tags generated from ~200 human and mouse tissues respectively suggests that the use of ATI sites is a common feature of protein coding genes and often results in modified N termini with potentially distinct functions (Carninci *et al*., 2006). In both human and mouse, these tags form approximately 200,000 tag clusters which map to ~35,000 loci and ~80% of known protein coding loci are covered by at least one tag cluster. When only protein coding genes were considered, 58% were found to make use of alternative promoters and 93% of these were predicted to result in the use of distinct start codons which for some genes occurred in a tissue specific manner. Hierarchical clustering of expression levels for all tag clusters revealed distinct global patterns of promoter usage associated with specific tissues, particularly lung, brain and liver.

Experiments which use both SAGE and CAGE have been proposed to allow independent profiling of the 5′ and 3′ ends of transcripts expressed in a single tissue sample (Wei *et al*., 2004). An interesting extension of the 32†profiling of SAGE and 52†profiling of CAGE described above has been reported as ‘gene identification signature’ (GIS) analysis (Ng *et al*., 2005). This approach allows the simultaneous profiling of the 5′ and 3′ end of a transcript by generating paired-end-tags (PETs) from random cDNAs followed by tag concatenation and sequencing. The advantage of this method over combining SAGE and CAGE is that each PET sequence represents a linked start and end position from a single transcript rather than two independent pools of tags representing start and end positions.

‘Next generation’ sequencing methods

The emergence of ‘next generation’ massively parallel sequencing technologies (Metzker, 2005) has the potential to dramatically improve the utility of sequence based approaches for profiling transcript diversity. The parallel sequencing of many templates on a single compact array was first published in 2000 by a group at Lynx Therapeutics Inc. (Brenner *et al*., 2000). This approach, described as massively parallel signature sequencing (MPSS) involves the creation of an array of microbeads, each coupled to a single DNA template, which are used for a ligation-based sequencing protocol involving fluorescently labeled adaptors. Monitoring of fluorescent signals as the sequencing reaction progresses is accomplished by a charge-coupled device (CCD) detector and image analysis, resulting in the simultaneous generation of millions of short sequences

(16–20 bp). This entire process takes place in a flow cell with the array of microbeads remaining in a dense monolayer and reagents flowing past. The accuracy of this platform for profiling gene expression was assessed by generating ~1.6 million sequences from cDNAs derived from a human cell line and comparing these to EST sequences generated by conventional Sanger sequencing. The resulting qualitative comparison of the most highly expressed genes seemed promising but far from definitive and early MPSS experiments identified strong biases related to the GC content of expressed sequences (Siddiqui *et al*., 2006). Lynx Therapeutics Inc. has since been acquired by Solexa Inc. and their platform has been modified and is now generally described as sequencing by synthesis (SBS) rather than MPSS. Although the technology is now being increasingly deployed and has been used successfully to identify genomic mutations in cancer cells (Thomas *et al*., 2006) and profile small RNAs of a plant genome (Lu *et al*., 2005), publications describing its use to profile transcript diversity are lacking. A competing platform which may also be described as a highly parallel SBS approach has been developed by Roche/454 Life Sciences Inc. and is capable of producing ~200,000 reads of ~100 bases in length from a single run (Margulies *et al*., 2005; Leamon *et al*., 2007; Bainbridge *et al*., 2006). In this platform, SBS occurs on a fiber-optic slide with approximately 1.6 million wells (each 44 μm in diameter). The sequencing reaction itself is referred to as 'pyrosequencing', in which fluorescently labeled nucleotides are sequentially washed over the slide and incorporation of each base into a growing complementary strand of a single stranded template DNA is simultaneously observed for all wells by a CCD detector. Homopolymeric sequences in the template DNA result in the incorporation of multiple nucleotides in a single cycle and must be resolved by analyzing the magnitude of fluorescence for each well. Several groups have used this sequencing platform to sequence SAGE-like libraries consisting of tags representing transcript ends or PETs (Gowda *et al*., 2006; Ng *et al*., 2006; Nielsen *et al*., 2006). These experiments are conceptually similar to SAGE but are able to produce increased tag counts at reduced cost and have been found to produce gene expression estimates that are similar to long SAGE data ($R^2 = 0.96$) (Nielsen *et al*., 2006). This approach was recently used to profile the 5′ ends of Maize transcripts and exhibited a considerable potential for identifying alternate transcript initiation sites (Gowda *et al*., 2006). Similarly, a combination of PET library construction and 454 sequencing was used to generate over 450,000 PETs from the human breast cancer cell line, MCF7 (Ng *et al*., 2006). Of these, ~136,000 could be mapped unambiguously to ~21,000 unique loci, and 25% of these represented candidate novel alternative transcript initiation sites or alternative polyadenylation sites. Finally, a recent experiment described the use of Roche/454 sequencing to profile full-length transcripts expressed in polyA+ purified RNA from the LnCAP prostate cancer cell line (Bainbridge *et al*., 2006). This direct sequencing of full-length transcripts avoids the artifacts associated with library construction and cloning and does not limit the resulting ESTs to the ends of transcripts. The approach was successful in identifying 25 novel AS events involving known exons but the short read lengths (average of ~100 bp), overrepresentation of a small number of highly expressed genes, and unexpected bias towards transcript ends limited the number of reads which were informative of splice site selection.

Limitations of sequence based approaches

Only 3–5% of the transcripts in a cell are mRNA molecules, with the remaining transcripts representing a few highly expressed ribosomal RNA (rRNA) species. The majority of rRNA transcripts can be removed by positively selecting for polyA+ sequences or less efficiently by filtering out rRNA species. However, even amongst the remaining mRNA transcripts, it is estimated that ~55% of these are redundant copies of the same mRNAs derived from only 4% of all protein coding loci (Alberts *et al*., 1994). Thus, even if a large number of tags can be produced efficiently, sequence based approaches are still faced with the problem of sequencing many transcripts from a few loci at the cost of failing to sample many other loci. For example in a recent test of Roche/454 Life Sciences GS20 sequencing for the profiling of a cDNA library, we found that ~110,000 reads could be mapped unambiguously to ~8,000 EnsEMBL loci but 39% of these corresponded to only 20 loci (Bainbridge *et al*., 2006). In addition to this issue of transcript redundancy, because of the complexity of mammalian biology, creating even a snapshot of the human transcriptome remains a daunting challenge. Assuming an average transcript size of ~2000 bp and an average of 300-500k transcripts per cell, complete profiling of a single cell type representing just one of hundreds or thousands of possible cell types would require 1 billion bp of sequence (the amount currently produced by a single run of the Solexa 1G device at launch specifications) (Ruan *et al*., 2004). Continued improvements in SBS technologies should be able to overcome these sampling limitations in the near future and will prove invaluable in characterizing even infrequently expressed transcripts.

Tools like the UCSC (Kuhn *et al*., 2006) and EnsEMBL (Hubbard *et al*., 2005) genome browsers can readily combine data from multiple sequence based methods by mapping them to a single reference genome sequence. Each of the sequencing technologies described can thus be used in a complementary fashion for the complex task of annotating the expression of all gene loci and act as a supplement to gene prediction algorithms which are based solely on the genomic sequence itself. Nevertheless, despite the recent advances in sequence based approaches and corresponding computational methods for profiling transcript diversity, other methods, particularly those based on microarrays remain a popular alternative for profiling a particular cell type and comparing expression across samples.

Microarray methods

Microarrays consisting of spotted cDNAs or short (25 to 60-mer) oligonucleotides have been used extensively to rapidly and simultaneously determine the overall level of mRNA expression of thousands of genes in a single sample. Briefly, a microarray is a small ordered grid of 'spots' (probes) each consisting of many copies of a single-stranded DNA sequence complementary to a small portion of a target gene. A microarray experiment involves extracting RNA from cells, converting the RNA to cDNA, labeling the cDNA molecules with a fluorescent dye, and hybridizing the labeled sample to an array. Each probe spot forms hybrids with copies of its target sequence and the degree of hybridization is measured by scanning the array and recording fluorescence

intensities. The magnitude of the intensity observed at each spot is thus a representation of the amount of probe/target hybridization and therefore an estimate of the number of copies of each target in the sample. Each probe on the array acts as a quantitative detector for a particular RNA sequence. Choosing the size and position of the sequence to target with each probe is an area of active development and largely determines the results of a microarray experiment. The general design and use of microarrays to detect gene expression has been reviewed extensively (Redkar *et al*., 2006) and each of the following microarray strategies are summarized in Table 1 and depicted in Plate IV.

'First generation' expression arrays

Despite the heavy use of microarrays for measuring gene expression, the use of these arrays to distinguish alternative transcripts has been limited. Spotted cDNA arrays use probes consisting of copies of entire cDNA transcripts or relatively large portions of them and are therefore unsuitable for the detection of alternative transcripts which have subtle differences involving only a small percentage of their total sequence content. Commercially available oligonucleotide microarrays such as those offered by Affymetrix Inc., NimbleGen Inc. and others are composed of sets of 10–20 short probe sequences per gene and therefore have higher resolution for detecting transcription (Plate IV). However, these designs and corresponding oligo d(T) based labeling procedures have heavily biased detection towards the 3′ end of transcripts (often confined to the UTR), limiting their ability to detect many alternative transcripts. Despite the limitations of these designs, the use of the raw probe values generated from these platforms to predict differential expression of alternate transcripts with variable exons at their 32†end has been described (Hu *et al*., 2001; Fan *et al*., 2006).

Whole genome and exon tiling arrays

Whole genome tiling arrays have emerged as a method of profiling transcription across large portions of the genome. These arrays consist of probes representing every non-repetitive base of a genome at 5–35 bp intervals (Plate IV). Because of this comprehensive approach, these arrays are not limited by the accuracy of gene annotations at the time of array design, but rather the completeness and accuracy of the genome sequence itself. Whole genome tiling arrays are theoretically capable of simultaneously determining the approximate exon-intron boundaries of all genes regardless of their current annotation status and also provide a quantitative measure of expression at every exon of every locus. Due to the size of the human genome, initial experiments focused on the smallest human chromosomes only (20, 21 and 22) (Kampa *et al*., 2004; Kapranov *et al*., 2002; Schadt *et al*., 2004). Arrays of 25- or 60-mer oligonucleotides were designed to tile across non-repetitive genomic sequence at 30–35 bp intervals and these arrays were hybridized with cytoplasmic polyA+ RNAs isolated from a variety of cell lines and tissues. These and subsequent experiments covering 10 human chromosomes at 5 bp resolution (Cheng *et al*., 2005) and the entire human genome at ~50 bp resolution (Bertone *et al*., 2004) have revealed considerable evidence for expression throughout the genome which has not been previously

annotated. Despite advances in microarray technology, the resources required to conduct such experiments are still daunting. For example, achieving ~50 bp resolution on both strands of the entire human genome required ~52 million probes distributed across 134 microarrays each of which was only hybridized with a single polyA+ RNA sample isolated from human liver tissue (Bertone *et al*., 2004). In other words, an extremely large number of probes were used to measure transcription of select regions of the genome (those that are actually expressed) from only a single tissue. Furthermore, despite the scale of this approach these arrays are unable to infer the connectivity of exons. Because of the comprehensive probe design strategy used in these arrays they are ideal for detecting novel gene, novel alternative exons within the introns of known genes and novel alternative exon boundaries. However, as the quality of gene annotation improves for the genome of interest, the value of using such a large number of speculative probes is reduced and space on the array can be reclaimed to be used more efficiently. Just as large scale sequencing efforts have revealed an unexpected level of transcript diversity at most loci, whole genome tiling array experiments have challenged accepted notions of what percentage of the genome is actually transcribed, indicating that it might be much larger than previously suspected (Johnson *et al*., 2005). Whole genome tiling arrays are likely to play an important role in continuing annotation efforts but currently have limited feasibility for profiling transcript diversity.

Affymetrix now offers exon tiling arrays which attempt to use array space more judiciously by designing probes for only those regions which are known to be expressed or predicted to be expressed by gene finding algorithms. Affymetrix's exon tiling arrays are created with a photolithographic *in situ* oligonucleotide synthesis platform and for human the design consists of a single array with ~5.5 million features corresponding to ~1.2 million known or predicted exons. This capacity allows each human exon to be covered by an average of 4 probes. This is by far the highest density array currently available but the oligo length is limited to 25-mers and medium to small scale custom designs are costly. The design strategy successfully overcomes some of the limitations of previous Affymetrix gene expression designs but these arrays are still unable to elucidate the connectivity of exons and may yield uninformative results when multiple isoforms are present in the same sample (Plate IV). Furthermore, Affymetrix currently only offers designs for the human, mouse and rat genomes. For researchers who do not wish to be limited to probes that only interrogate the exons of each gene or wish to study AT in additional species, a number of options are available for printed or bead based custom designs of up to ~150,000 features (Agilent, Illumina and others). The maskless photolithography procedure of NimbleGen remains the highest density custom array option with up to ~385,000 features and additional advantages such as the ability to create probes up to 60 nucleotides in length as well as variable length (isothermal) designs (Nuwaysir *et al*., 2002).

Splicing arrays

As discussed, 'traditional' microarrays have been designed to measure the expression of only a single canonical transcript of each gene and do not account for the existence

of alternate isoforms. The idea of using 'splicing' microarrays consisting of exon-junction and other probe configurations to detect AS events was first suggested by Douglas Black (Black, 2000). Since 2002, a number of groups have begun to experiment with measuring expression in the context of AT by using such modifications of existing microarray technology (Lee and Roy, 2004). ExonHit Therapeutics offers a commercial service for detection of AS in selected therapeutic targets (Lyddy, 2002; Mangasarian, 2005). Jivan Biologics offers the 'TransExpress™Whole Spliceome' array which includes probes for ~135,000 alternately spliced sites corresponding to ~23,000 human genes. The splice events selected for this array were identified by bioinformatic analysis of existing EST data. In addition to these commercial options, several groups have described the development of custom splicing arrays using commercially available *in situ* oligonucleotide synthesis or printing platforms.

A number of works have specifically addressed the theoretical and practical issues of designing custom splicing microarrays to detect AT events by conducting proof-of-principle experiments in a variety of metazoan species (Srinivasan *et al*., 2005; Castle *et al*., 2003; Clark *et al*., 2002; Johnson *et al*., 2003; Stolc *et al*., 2004; Pan *et al*., 2004). Issues addressed by these experiments include the following. (1) Accurately annotating gene models to assist in the selection of oligonucleotides. This involves the identification of all exons for every gene, the precise boundaries of each exon, and the putative connections of these exons. The utility of a splicing microarray is fundamentally limited by the accuracy and comprehensiveness of this annotation process. Defining exon regions as either 'constitutive' or 'alternative' by examining existing expression data is also desirable to facilitate within-gene normalization during analysis, (2) Storing gene models and annotations of splicing events in a computer interpretable format such as "splicing graphs" (Heber *et al*., 2002), (3) Selecting the number and types of AS events to profile. For example, one may wish to target only sequences within exon boundaries. If the identification of complicated splicing patterns is desired it may be prudent to target exon boundaries, exon junctions, and introns as well (Plate IV). Each of the array design strategies used to date falls into one of two general categories. In one case, transcript annotations based on existing expression data (ESTs, cDNAs, etc) are assumed to be an acceptable representation of the transcript diversity in the genome and used to identify known AT events which are then specifically targeted by the array. In the second case, the array design attempts to comprehensively profile all exons and splicing events regardless of existing expression evidence. This approach requires considerably more probes but it has the potential to identify the expression of novel transcription events, (4) Optimizing the specificity and thermodynamic properties of probes to improve the ability of each probe to accurately and reliably predict the presence of their target during hybridization. A uniformity of probe melting temperature (Tm) and length across the array is desirable. Furthermore, probes that form secondary structures, have low-complexity regions, match repetitive elements, or correspond to expressed sequences from multiple regions of the genome should be avoided. For members of gene families or genes with pseudogenes, it may not be possible to select specific probes. Furthermore, when targeting large exons and introns, choosing an 'optimal' probe is often straightforward, but when the target

sequence is constrained to a small exon or a specific exon junction or boundary this may not be possible, (5) Reducing 'half-junction crosstalk'. This term refers to a problem related to the use of exon junction probes such that each probe hybridizes over each half of its length to targets containing the same exon sequences in combinations other than that specifically targeted by the junction probe. For example, a probe designed to detect the juxtaposition of exon 1 with exon 3 (e1^e3) will hybridize on each half to RNAs containing e1^e2 and e2^e3. This crosstalk effect increases as the length of a probe is increased or hybridization stringency is reduced. The junction probe length that maximizes sensitivity and specificity has been empirically determined as 35–45 nucleotides in length (Castle *et al*., 2003; Srinivasan *et al*., 2005; Clark *et al*., 2002). Crosstalk can theoretically be reduced by offsetting the probe position on the exon junction or allowing the two halves to differ in length such that the difference in Tm between the two halves is minimized. The proof-of-principle experiments which have helped to resolve these five issues provide invaluable guidance for researchers wishing to create custom splicing arrays without spending considerable time and resources conducting optimization experiments. Furthermore, their results provide general evidence that the splicing microarray approach is feasible. For example, experiments using samples spiked with different mixtures of cloned human and *Drosophila* transcripts showed that an alternate isoform making up as little as 20% of a mixture of two isoforms could be detected by junction probes (observed fold differences were highly correlated with expected values over a range of 0.25 to 12) (Castle *et al*., 2003; Fehlbaum *et al*., 2005).

The first two large scale experiments with splicing microarrays were conducted in human, mouse and *Drosophila* (Johnson *et al*., 2003; Pan *et al*., 2004; Stolc *et al*., 2004). Johnson *et al*. (2003) conducted a genome-wide survey of AS in 52 human tissues using a total of 125,000 exon junction probes corresponding to the expected canonical junctions of 10,000 multi-exon genes. The authors observed that similar tissues tend to have similar AS patterns and cell lines have their own distinct patterns, in particular being characterized by the expression of fewer genes but more variants of those genes. By extrapolating from their results and comparing to EST data the authors predicted that 74% of all human genes are alternatively spliced. A similar approach was used to analyze ~3,000 previously observed AS events in 10 mouse tissues (Pan *et al*., 2004). Based on RT-PCR validations of the predictions of their splicing microarray the authors determined that the array could predict differential expression of isoforms between tissues with a specificity of approximately 80%. The data described in this initial experiment has recently been analyzed to show that exons that have varying expression levels across mouse tissues are more likely to be a multiple of 3 in length (perhaps indicating a selection for maintenance of reading frame) and are highly conserved relative to constitutively spliced exons (Xing and Lee, 2005). This data has also been used to investigate the potential coupling of AS and nonsense-mediated mRNA decay as a global means of controlling transcript abundance (Pan *et al*., 2006).

As the number of published splicing microarray experiments has increased, the variety of analysis methods has also increased (Cuperlovic-Culf *et al*., 2006). Nevertheless, the availability of suitable analysis methods with open source software

implementations remains a challenge to researchers who wish to conduct their own splicing microarray experiments. Standard methods for normalization, background correction and summarizing multiple probe values into a single gene- or exon-level expression estimate may be used (Butte, 2002; Bolstad *et al*., 2003; Irizarry *et al*., 2003) but methods which specifically address the identification of differences at the level of alternative transcripts are still required. To date, at least seven distinct analytical methods for identifying differences in isoform expression from splicing microarray data have been described: (1) Splicing index values (Clark *et al*., 2002; Srinivasan *et al*., 2005; Li *et al*., 2006), (2) ASAP (Le *et al*., 2004), (3) splice and neighborhood algorithms (Fan *et al*., 2006; Hu *et al*., 2001), (4) analysis of splice variation (ANOSVA) (Cline *et al*., 2005), (5) sequence based splice variant deconvolution (Wang *et al*., 2003), (6) GeneASAP (Shai *et al*., 2006), and (7) MIDAS (www.affymetrix.com). Although each of these methods uses different mathematical and statistical techniques, the general goal of each is to identify alternative exons, junctions, or whole transcripts that are differentially expressed between two samples. Identifying such events invariably involves some attempt to correct for changes in expression at the gene level. For example, the use of a simple 'splicing index' calculation was proposed to identify AS events (Clark *et al*., 2002). A splicing index is determined by first comparing the expression of each exon to the expression value for the entire gene within a single sample. This results in a 'within-gene' normalized value for each exon which can then be compared across samples to create the splicing index. Statistical methods such as MIDAS also use within-gene normalized values but attempt to identify significant differentially spliced exons by considering the magnitude and variability of exon expression within grouped samples compared to across sample groups (e.g., ten normal versus ten cancer samples).

Using the developments in splicing microarray design and analysis described above, several research groups have applied these arrays to the study of specific biological problems. These include estimating the global prevalence of AT in tissues and throughout development, assessing the implications of AT for protein diversity, studying splicing regulation at the level of trans-acting factors, defining novel *cis*-acting splicing motifs, and identifying isoforms with disease relevance.

Relogio *et al*. (2004) published the first experiment using microarray technology to specifically address the role of AS in a cancer model. This group designed a custom array to measure the expression of 86 splicing-related genes and known splicing events in 10 cancer genes and applied their array to RNAs derived from four cell lines representing different stages of Hodgkin lymphoma tumors. Clustering of the microarray results for 100 splicing events revealed distinct patterns for each of the four tumor stages. A similar study used splicing microarrays to identify differential expression of alternate transcripts between estrogen receptor positive and negative breast cancer cell lines (Li *et al*., 2006). Zhang *et al*. (2006) predicted that profiling expression at the level of individual exons and AT events with a splicing microarray would improve the accuracy of expression based cancer classification compared to using overall mRNA expression levels. They demonstrated this by conducting a classification of 38 cancer and normal prostate tissues by measuring the expression of 464 isoforms of ~200

genes and concluded that profiling the expression of alternative transcripts increased the information content by at least 30% compared to conventional microarray data. In addition to studying human disease, splicing microarrays have also been shown to have great potential for defining a global 'splicing code' by studying the expression of thousands of exons and identifying novel sequence motifs as well as how the arrangement of these motifs and their interaction with particular *trans*-acting factors influences the splicing of specific exons in a tissue dependent manor (Sugnet *et al*., 2006; Ule *et al*., 2006; Blanchette *et al*., 2005; Ule *et al*., 2005). For example, one group used a splicing microarray to study the global effects of RNAi knockdowns of four splicing regulators (two hnRNPs and two SR proteins) (Blanchette *et al*., 2005). Knocking down each of these four proteins affected a variable number of splicing events, ranging from ~50 to more than 300. Since their array design was limited to only those events that had been previously observed (~8,000 events observed for ~3,000 genes in EST/mRNA data), these are likely to be underestimates.

We have reviewed a number of preliminary experiments describing the creation and use of custom splicing microarrays and commercially available solutions have recently emerged for the human, mouse and rat genomes. Unfortunately there are currently no simple solutions available for the researcher who wishes to create a custom splicing array for another species, subset of genes or design philosophy not represented by these products. Although considerable optimization has been reported and methods for creating such designs and analyzing the resulting data have been published by several groups, actually implementing these experiments remains time consuming. Continued expansion of commercially available solutions as well as the creation of open source design and analysis tools are likely in the near future and will accelerate the adoption of splicing microarrays as a tool for transcriptome analysis.

Given the advantages and disadvantages (Table 1) of both the sequence- and microarray-based approaches for profiling transcript diversity, several groups have begun to combine complementary computational and experimental approaches and thereby develop a more comprehensive view of the mammalian transcriptome (Gustincich *et al*., 2006). As a result of these efforts and the continued compilation and synthesis of disparate genome scale expression data sets in genome browsers such as the UCSC (Kuhn *et al*., 2006) and EnsEMBL (Hubbard *et al*., 2005) browsers, many researchers now have access to a highly detailed survey of the diversity of transcripts generated by their genes of interest.

Validation and visualization

Due to methodological advances and increases in information as we have described, researchers are now increasingly able to identify the complex pattern of alternative transcripts generated by the genes under study in their laboratory. It is therefore becoming increasingly important to have a wide range of tools and protocols for (1) the *in silico* visualization of transcript diversity for a gene of interest (Table 2); (2) the visualization of the expression of particular isoforms in cell lines or *in vivo* models; and (3) determining the function of specific isoforms.

Determining the relative mRNA expression of known or predicted isoforms of a single gene in a tissue of interest is typically accomplished by Northern blot analysis or by semiquantitative or quantitative RT-PCR. Similarly, protein-level expression of isoforms with significantly different sizes can be confirmed by SDS-PAGE and Western blot analysis with an antibody that recognizes a constitutive portion of the gene. Visualizing the spatial expression of isoforms at the mRNA level can be accomplished by *in situ* hybridization with digoxigenin labeled riboprobes specific to each isoform (David *et al*., 2002). Visualizing spatial expression of isoforms at the protein level by immunohistochemistry is limited by the availability of antibodies specific to the isoforms of interest and the labor-intensive, time-consuming nature of raising novel antibodies to specific isoforms. Although databases of antibodies have been described, considerable effort may still be required to determine which, if any available antibodies will distinguish between the isoforms of interest (Major *et al*., 2006). *In vivo* methods of visualizing alternate isoforms have been described for model organisms such as *C. elegans* (Kuroyanagi *et al*., 2006) and mouse (Kemp *et al*., 2005). In general, the visualization experiments described here are labor intensive and difficult to apply to a large number of isoforms.

Functional validations of the effects of particular isoforms can be studied in a number of ways. To date, most studies have attempted to simply infer the function of isoforms by observing differences in expression level, subcellular localization, post-translational modifications and other modifications in cells where the gene of interest is thought to play some role (Nanjundan *et al*., 2006; Vegran *et al*., 2006). Examples of direct manipulation of the expression of an isoform are less common. In many cases an RNA interference based approach should be able to specifically 'knock down' an isoform of interest in cell culture and considerable resources already exist to facilitate these kinds of experiments (Paddison *et al*., 2004). Over-expression of a single canonical isoform in an expression vector which has been transfected into a suitable cell line is already common place. Similarly creation of transgenic mice expressing a particular isoform has been widely reported (by definition selecting a single isoform is required). Altering expression of an isoform can be used in conjunction with studies of particular functions of interest such as apoptosis or cell survival assays. Differences in the protein-protein interactions of alternate isoforms can be studied by accepted methods such as co-immunoprecipitation of expected partners or immunoprecipitation of tagged isoforms followed by HPLC-MS to identify interacting partners (Figeys *et al*., 2001). Studying multiple isoforms in these kinds of experiments, although more labor intensive will become increasingly common as researchers become aware of the transcriptional diversity generated by genes of interest.

Functional significance of alternative transcription

As large scale experimental and bioinformatic approaches have begun to identify the diversity of transcription across the genomes of several species, parallel efforts to study the functional significance of this diversity have also been reported. One area of intense debate has been the effort to estimate the proportion of AT events that are functional compared to that which represents 'transcription noise'. Other areas which

have generated a large number of publications include the effort to identify general themes by which AT influences cellular biology, the study of particular functional classes of genes that are affected by AT and its potential role as a means of globally regulating gene expression. Finally, the implications of AT for the study of human disease has received increasing attention in recent years. For example, the emergence of a 'transcription code' has implications for the identification of potential disease mutations; increased knowledge of transcriptome complexity will influence strategies for identifying therapeutic targets; and the mechanisms of RNA processing itself are being considered as a means of directly modulating disease states.

How much alternative transcription is functional?

Although the notion that transcript diversity is more prevalent than originally thought is generally accepted, the percentage of alternative transcripts with biologically relevant functions remains a topic of debate. Detailed studies of single genes or pathways have identified differing functions for alternate isoforms. Although these single gene studies hint at the mechanisms by which AT allows a diversity of functions to be encoded from a single locus, they do not confirm the role of AT as a global means of generating biologically relevant diversity in the proteome. To address this outstanding question, a number of studies have attempted to use conservation of AT events between species to infer the fraction of all events that are functionally significant as opposed to transcription 'noise' caused by random splicing errors or observations of immature transcripts derived from the nucleus. The resulting estimates for the percentage of alternative events represented in EST data that are conserved between human and mouse range from 11 to 61% depending on the study. For example, to estimate the subset of alternatively spliced exons that are functional, one group used ESTs to identify exon skipping events which occur in both humans and mouse (Sorek and Ast, 2003). Of a total of 980 exons identified as alternatively skipped in humans, 25% were also skipped in mouse. The authors of this study presented further evidence that the characteristics of the conserved subset of alternate exons were distinct from those of the non-conserved exons and suggested that the majority of non-conserved events are non-functional. A similar study observed AS events in 2,603 human genes and their mouse orthologs (Pan *et al*., 2004). The authors found that of all the orthologous exons that are alternatively spliced in human or mouse, 16% are alternatively spliced in both species, and the remaining 84% represent species-specific events. By considering events represented in multiple transcripts from multiple tissues for both human and mouse, the authors estimated that at least 24% of these events represent true examples of species-specific AS. It has also been argued that studies which utilize EST data will underestimate the conservation of AS between mouse and human because they rely heavily on the level of transcript coverage (Thanaraj *et al*., 2003). In other words, conservation of a splicing event observed in human is often not observed in mouse simply because the EST sampling depth is too low and by chance it has not been observed. These authors conducted a conservation study similar to those previously described but also developed a statistical model to estimate the 'true' level of conservation by extrapolating from existing levels

of transcript support. Using this model, they estimated that 61% of alternatively spliced junctions are conserved between mouse and human. In contrast, a more recent study found that only 11% of the alternatively spliced exons in humans are conserved in mouse and suggested that the majority of AS events seen in EST/cDNA data represent aberrant splicing, disease-specific splicing or events that are functionally relevant but specific to humans (Yeo *etal*., 2005). One theme that emerges from these works is the considerable disagreement in the literature as to what percentage of AT is truly conserved and indeed what percentage of non-conserved events might be functional but species-specific events that emerged since the divergence of human and mouse 85 million years ago. AT events that are not conserved between human and mouse tend to be expressed at lower levels and may serve as an evolutionary mechanism for testing novel proteins without disrupting the function of the canonical isoform and interfering with the normal functions of the cell (Pan *et al*., 2005; Pan *et al*., 2004). The 'lesser' form is thus unlikely to be detrimental, is relatively free of constraints, can evolve rapidly and in some cases gain a function that is driven by positive selective pressure. It has been suggested that incorporation of novel exons or boundaries in this way represents a major form of gene evolution which is distinct from evolution by gene duplication. This hypothesis is based on the observation that genes which are part of gene families that have arisen by duplication generally have few alternate transcripts, whereas 'singleton' genes have high rates of AT (Kopelman *et al*., 2005; Su *et al*., 2006).

It is reasonable to assume that most conserved AT events are functional, that some as yet unknown fraction of non-conserved events are also functional and the remaining fraction are not functional. Although the percentage of events falling into each of these categories remains an area of active debate, any study of AT will certainly be complicated by some level of expression 'noise' with limited functional relevance.

How does alternative transcription influence the proteome?

The number of AT events that result in a protein with a modified biological function is currently a topic of debate. The concept that this subset of AT events could increase the functional diversity of the human genome by generating a combinatorial output of proteins from a genome of perhaps less than 30,000 genes has gained acceptance in recent years (Black, 2000; Maniatis and Tasic, 2002; Roberts and Smith, 2002). Furthermore, AT of specific genes has been shown to regulate transcript abundance via nonsense mediated decay, alter the subcellular localization of proteins, influence enzymatic activity, modify protein stability, and alter posttranslational modifications (Stamm *et al*., 2005). One of the most striking examples of AT producing diverse products from a single gene locus was observed for the DSCAM gene of the model organism *Drosophila melanogaster* (Schmucker *et al*., 2000). When transcribed, this gene selects exons from a set of mutually exclusive alternate exons at four positions. Specifically, exons 4, 6, 9 and 17 in each transcript are selected from 12, 48, 33, and 2 possible alternatives respectively. This remarkable arrangement is capable of producing 38,016 possible unique DSCAM transcripts. Cloning and sequencing a sample of 50

random cDNAs for this gene yielded 49 unique transcripts which result in distinct proteins with differing abilities to form neuronal connections. A comparably dramatic level of diversity was recently described for the human basonuclin 2 (BN2) locus, a zinc finger protein which is expressed ubiquitously and thought to function in RNA processing (Vanhoutteghem and Djian, 2006). All 23 exons of this gene are alternative and each transcript independently uses one of six promoters and four polyadenylation sites. To date more than 100 distinct BN2 mRNA isoforms have been produced, but a staggering ~90,000 are possible.

AT may result in the production of protein isoforms that are functionally distinct in a number of ways. It has been suggested that this diversity is realized in part through alterations in protein-protein interactions. Specific examples of genes such as SMRT which produces isoforms differing in their interaction with thyroid hormone receptors have been studied in detail (Goodson *et al*., 2005). Furthermore, global analysis of EST data has shown that AT events disproportionately affect domains involved in protein-protein interactions (Resch *et al*., 2004). Although only 10% of AS events can be shown to completely remove or insert a known functional domain, many of the remaining 90% of AS events are predicted to affect loop structures in proteins which are thought to mediate protein-protein interactions (Wang *et al*., 2005). A recent study also found that the majority of changes observed in isoforms do not affect complete protein domains and based on an analysis of the 3D structures of alternative isoforms concluded that AT modulates the activity of protein networks and associated signaling pathways indirectly by altering the structural core and resulting stability of proteins (Yura *et al*., 2006). For example, replacing a stable domain with an unstable domain in a protein could alter the spatial orientation of other domains resulting in a protein with a distinct conformation and affinity for interaction partners. These observations have led to the general speculation that AT outcomes profoundly influence the protein interaction network of a cell. Supporting this hypothesis is the observation that genes with large numbers of isoforms tend to have many interactions and represent central nodes in protein-protein interaction networks (Hughes and Friedman, 2005). In addition to modifying protein interactions, another common effect of AT is the modification of subcellular localization in which alternative isoforms differ in their signal peptides and/or transmembrane domains (Davis *et al*., 2006; Xing *et al*., 2003). Such modifications can result in post-translational transport to different cellular compartments or the production of a soluble protein rather than a membrane bound one.

As discussed, AT can presumably influence protein interactions, protein stability and subcellular localization and through each of these types of effects has the potential to influence signaling pathways. These observations suggest some of the general modes by which AT influences the function of any protein. Efforts to identify whether genes of particular functional classes are more likely to be modulated by AT have also been reported. For example, Takeda *et al*. (2006) used a comprehensive analysis of 55,000 cDNAs to determine that the gene classes (according to Gene Ontology terms) which are most affected by AT are: nucleic acid binding, transcription factor activity, DNA-binding, protein tyrosine kinase activity, transporter activity, zinc ion binding, insulin-like growth factor-binding, ATP binding, catalytic activity, and oxidoreductase activity.

Analysis of cDNA, EST and MPSS data in mouse found that 75% of all kinases and phosphatases have alternate isoforms and analysis of these variants revealed several tethered and soluble, secreted isoforms which were predicted to be catalytically inactive and therefore might act as dominant negative forms by competing with other isoforms for ligands and substrates (Forrest *et al*., 2006). Similar studies have demonstrated the prevalence of functional isoforms within the G protein coupled receptor family (Bjarnadottir *et al*., 2006), zinc-finger-containing proteins (Ravasi *et al*., 2003) and apoptosis genes (Schwerk and Schulze-Osthoff, 2005).

Finally, it is important to note that production of a transcript variant which does not seem to produce a functionally distinct protein may still have functional consequences for the cell by altering the level of gene expression. For example, AS is speculated to act as a gene expression 'switch' whereby genes are effectively turned off by changes in the expression of a splicing factor which disrupts their normal splicing and silences their expression by triggering nonsense mediated decay (NMD). NMD targets transcripts with 'premature termination codons' which are recognized by the transcription machinery and degraded rather than producing a potentially detrimental protein product. In this system, transcription of a gene may still occur at the same rate but since the mRNA products are quickly degraded the gene's function is essentially silenced. Recent studies have suggested that coupling of NMD and AS is an important but overlooked mechanism of regulating gene expression (Hillman *et al*., 2004; Lewis *et al*., 2003; Cuccurese *et al*., 2005). In addition to NMD which is triggered by events within the coding region of a transcript, AT within UTRs may also act as a global means of controlling gene expression by altering mRNA stability and translational efficiency in a tissue specific manner (Hughes, 2006). In this case a valid mRNA is produced and would seem to result in production of a normal protein but due to sequence modifications outside the coding region, the stability of the transcript or its rate of translation is modified.

What are the implications of transcript diversity for the study of human disease?

The role of AT in human disease has received increasing attention in recent years (Caceres and Kornblihtt, 2002; Faustino and Cooper, 2003; Garcia-Blanco *et al*., 2004; Stoilov *et al*., 2002). In particular, the apparent existence of a defined 'transcription code' has implications for the identification of potential disease causing genomic variants (e.g., point mutations, insertions, deletions). This code can be considered as the combination of (1) regulatory sequence motifs of a transcribed region, and (2) RNA and protein factors which comprise the machinery responsible for correct transcription initiation, splicing and polyadenylation. Genetic changes that have the potential to alter normal transcription and contribute to human disease can thus be classified into two groups, '*cis*-acting' variants which affect sequence motifs within each gene locus and '*trans*-acting' variants which affect components of the transcription machinery itself. Examples of human disease involving both of these classes of variants have been reviewed in the context of neurological disorders and cancer (Srebrow and Kornblihtt, 2006; Licatalosi and Darnell, 2006).

Disease associated transcripts may arise by the occurrence of *cis*-acting mutations within the transcription regulatory elements of a single gene (Plate I) and many examples of heritable diseases have been shown to result from point mutations leading to aberrant splicing of a gene. Such mutations may result in aberrant skipping of a canonical isoform, inclusion of a 'cryptic' exon that is not normally used or simply an alteration of the ratio of alternative isoforms normally expressed (Pagani and Baralle, 2004). According to the Human Gene Mutation Database, ~10% of all disease associated mutations involve splice sites (Stenson *et al*., 2003). In addition to splice site mutations, many other mutations may affect splicing regulatory sequences such as exonic and intronic splicing enhancers and silencers (Cartegni *et al*., 2002). For example, analysis of the effects of mutations in the well studied human disease genes ATM (ataxia-telengiectasia, OMIM #208900) and NF1 (Neurofibromatosis type I, OMIM #162200) suggests that as much as 50% of all exonic mutations, silent or otherwise exert their influence by causing splicing defects (Ars *et al*., 2000; Teraoka *et al*., 1999). Many of these mutations are at splicing regulatory sites, not the actual splice sites. Until recently the only mutations associated with disease that were predicted to affect splicing of a gene product were those associated with the splice acceptor and donor sites specifically. Increasing knowledge of the additional motifs which influence AT has expanded the number of mutations which are predicted to affect transcription. Many non-synonymous mutations may have a more pronounced effect than causing a single amino acid change and in fact may influence the inclusion or exclusion of entire exons. Similarly, many synonymous mutations or mutations outside of the coding sequence which might seem to be non-functional may also influence exon content. A number of studies have recently begun to investigate the effects of mutations in known disease genes at positions other than the actual splice sites and preliminary attempts to predict and validate the effect of point mutations on AS in splicing regulatory motifs such as ESEs have been reported (Cartegni and Krainer, 2002; Wang *et al*., 2004; Zhang and Chasin, 2004; Smith *et al*., 2006). Some of these studies rely on the observation of mutations and their effect on the splicing of specific genes. Others attempt to computationally predict the effect of mutations occurring within exons or introns on the splicing outcome of a gene. Similar efforts are needed to understand the true implication of mutations on the use of alternate transcription initiation sites and polyadenylation sites. In other words, although it has long been accepted that polymorphisms or mutations affecting 'regulatory' sequences may affect the tissue- or developmental-specific expression level of a gene, it is now becoming clear that an entirely additional set of 'regulatory' changes act by influencing AT without necessarily changing the level of expression.

Reports documenting disease associated mutations that occur in *trans*-acting factors of the splicing machinery and that result in the aberrant processing of several genes are less common than those involving *cis*-acting mutations but a few examples are well documented. Two forms of the familial disease Retinitis pigmentosa, RP18 and RP13 are caused by mutations in precursor mRNA processing factors 3 and 8 respectively (OMIM #601414 and #600059). For some diseases associated with aberrant splicing such as certain cancers, it is often not known whether a cancer-associated AT event arises because of acquired or inherited mutations in *cis*-acting transcription regulatory

motifs or changes in the expression of *trans*-acting splicing factors. However, in some cancers such as chronic myeloid leukemia (CML) the evidence for involvement of splicing factors is becoming more convincing. For example, the Bcr-Abl fusion product of CML has been shown to cause changes in the expression of genes involved in pre-mRNA splicing resulting in the aberrant splicing of a cascade of other genes which in turn contributes to pathogenesis (Salesse *et al*., 2004). Bcr-Abl dependent over-expression of the splicing gene SR Protein Kinase 1 (SRPK1) was observed in CD34+ blood cells and this over-expression was associated with aberrant splicing of apoptosis and differentiation genes such as Pyk2, SLP65, BTK and Ikaros. Furthermore, both the expression of Bcr-Abl and the aberrant splicing of Pyk2 were partially reversed by treatment with the kinase inhibitor STI571 (Imatinib/Gleevec®). Since these studies, other cancer causing fusion proteins in leukemia and Ewing Sarcoma have also been hypothesized to contribute to aberrant splicing by affecting the expression of splicing factors.

Regardless of whether the effect is via a *cis*- or *trans*-acting effect, the general potential for splice variants to act as diagnostic or prognostic markers or novel therapeutic targets for complex diseases such as cancer seems promising (Brinkman, 2004). The observation that the genome is capable of producing a dramatic diversity of products from a relatively small number of loci has already begun to influence strategies for identifying therapeutic targets. For example, a number of studies have used bioinformatic approaches to identify cancer-specific splice variants by analyzing the content of human EST, SAGE and microarray repositories (Hui *et al*., 2004; Kirschbaum-Slager *et al*., 2005; Xu and Lee, 2003; Kirschbaum-Slager *et al*., 2004). Increasing the resolution of gene expression screens for therapeutic targets to profile individual exons and AT events has the potential to identify previously unobserved and potentially more definitive events specific to disease states. For example the application of exon tiling and splicing microarrays or deep sequencing with SBS platforms to the comparison of normal versus diseased tissues, drug responders versus non-responders and other relevant comparisons seems certain to yield novel biomarkers which would have been previously impractical to detect. Since AT can create functionally significant variants, searching for these variants in target discovery efforts should result in the identification of distinct protein isoforms associated with disease which may be more useful targets than proteins that are simply up- or down-regulated in disease. For example the Bcl-x gene is alternatively spliced to form a long isoform which is anti-apoptotic (Bcl-x_L) and a short isoform which is pro-apoptotic (Bcl-x_S) and targeting this locus by inactivating one isoform or simply shifting the ratio of isoforms has been proposed as a cancer treatment (Mangasarian, 2005). Many targets may have evaded detection in previous gene-expression studies of disease because of a technological inability to profile this kind of transcript diversity from each locus. The identification of targets for the development of small molecule drugs and therapeutic antibodies (Wiles and Andreassen, 2006) will thus be greatly enhanced by considering alternate isoforms and their subtle differences in amino acid content. In addition to the identification of drug targets, AT also has implications for pharmacogenomics and there is evidence that polymorphisms which alter splicing may underlie differences

in drug efficacy and toxicity between patients. For example, the most common polymorphism of CYP2D6, a gene which is responsible for the metabolism of at least 40 drugs, results in the aberrant splicing and production of a non-functional protein from this gene (Bracco and Kearsey, 2003).

Targeting specific isoforms with small molecule or antibody therapies is a simple extension of current drug design efforts but targeting the transcriptional machinery itself has also been proposed as a means of altering gene expression and treating disease. Proof-of-principle experiments describing the screening of drugs that target splicing factors such as SR-proteins to inhibit aberrant splicing or produce a desired splicing outcome have been reported (Yeo, 2005). Antisense oligonucleotide therapies to directly manipulate the splicing patterns of specific disease genes have also been described (Wilton and Fletcher, 2005). These molecules can be used to influence splicing in many ways such as preventing the inclusion of an aberrant exon by masking a cryptic splice site, or forcing an exon-skipping event to allow nonsense or frameshift mutations to be by-passed. Current studies have only begun to address the ways in which an understanding of AT can influence the study of human disease by enhancing the identification of therapeutic targets, allowing the design of novel types of therapies and predicting the efficacy and toxicity of drugs for individual patients.

Conclusions and future perspectives

Since the completion of the human genome sequence, a focus of genome research has been the study of the transcriptome, particularly the identification and annotation of genes. Efforts to profile the expression of genes across tissues and developmental stages, identify the regulatory elements which control gene expression and characterize the genomic variations between individuals which influence these patterns have been widely reported. The phenomenon of AT described in this chapter dramatically increases the complexity of the transcriptome and the functional diversity of the proteome and therefore has profound implications for biology. We have described past and present methods for studying transcript diversity and highlighted recent advances in sequencing and microarray technology which will make more detailed analyses of the transcriptome possible in the future. Specifically, the advent of high-throughput sequence-by-synthesis approaches and the increased density and reduced cost of microarrays will allow more comprehensive and quantitative studies to be conducted. These advancements will also allow current biological challenges to be addressed. For example, understanding the regulation of AT will require extensive genome-wide analysis of the complex interplay of hundreds of *trans*-acting factors as well as the sequence composition and configuration of the *cis*-acting regulatory sequences they interact with. Other important challenges include determining the relevance of AT induced NMD as a means of globally regulating gene expression; the potential roles of AT in gene evolution; the functional significance of both conserved and non-conserved AT events; and the amount of natural variation in AT between individuals. In addition to these broad challenges, advances in transcriptome profiling will have a positive impact on the identification of developmental, tissue and disease specific alternative transcripts. The increased

resolution and sophistication of splicing microarrays and increased sampling depth of SBS approaches will result in improved disease classification as well as accelerated identification of novel therapeutic targets and disease markers.

REFERENCES

Adams M.D., Kerlavage A.R., Fields C., Venter J.C. 1993. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* **4:**256–267.

Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J.D. 1994. *Molecular Biology of the Cell*. Garland Publishing Inc., New York.

Ars E., Serra E., Garcia J., Kruyer H., Gaona A., Lazaro C., Estivill X. 2000. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet*. **9:**237–247.

Bainbridge M.N., Warren R.L., Hirst M., Romanuik T., Zeng T., Go, A., Delaney A., Griffith M., Hickenbotham M., Magarini V., Mardis E., Sadar M., Siddiqui A., Marra M., Jones S. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7:**1–11

Baross A., Butterfield Y.S., Coughlin S.M., Zeng T., Griffith M. 2004. Systematic recovery and analysis of full-ORF human cDNA clones. *Genome Res.* **14:**2083–2092.

Berget S.M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem*. **270:**2411–2414.

Bertone P., Stolc V., Royce T.E., Rozowsky J.S., Urban A.E., Zhu X., Rinn J.L., Tongprasit W., Samanta M., Weissman S., Gerstein M., Snyder M. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:**2242–2246.

Bjarnadottir T.K., Geirardsdottir K., Ingemansson M., Mirza M.A., Fredriksson R., Schioth H.B. 2007. Identification of novel splice variants of Adhesion G protein-coupled receptors. *Gene.* **387:**38–48.

Black D.L. 2000. Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* **103:**367–370.

Black D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem*. **72:**291–336.

Blanchette M., Green R.E., Brenner S.E., Rio D.C. 2005. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. *Genes Dev*. **19:**1306–1314.

Boguski M.S., Lowe T.M., Tolstoshev C.M. 1993. dbEST — database for "expressed sequence tags". *Nat. Genet*. **4:**332–333.

Bollina D., Lee B.T., TanT.W., Ranganathan S. 2006. ASGS: An alternative splicing graph web service. *Nucleic. Acids. Res.* **34:**W444–447.

Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:**185–193.

Bonaldo M.F., Lennon G., Soares M.B. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6:**791–806.

Boon K., Osorio E.C., Greenhut S.F., Schaefer C.F., Shoemaker J., Polyak K., Morin P.J., Buetow K.H., Strausberg R.L., De Souza S.J., Riggins G.J. 2002. An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. USA* **99:**11287–11292.

Bracco L., Kearsey J. 2003. The relevance of alternative RNA splicing to pharmacogenomics. *Trends Biotechnol*. **21:**346–353.

Brenner S., Johnson M., Bridgham J., Golda G., Lloyd D.H., Johnson D., Luo S., McCurdy S., Foy M., Ewan M., Roth R., George D., Eletr S., Albrecht G., Vermaas E., Williams S.R., Moon K., Burcham T., Pallas M., DuBridge R.B., Kirchner J., Fearon K., Mao J., Corcoran K. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol*. **18:**630–634.

Brett D., Pospisil H., Valcarcel J., Reich J., Bork P. 2002. Alternative splicing and genome complexity. *Nat. Genet*. **30:**29–30.

Brinkman B.M. 2004. Splice variants as cancer biomarkers. *Clin. Biochem*. **37:**584–594.

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

Butterfield Y.S., Marra M.A., Asano J.K., Chan S.Y., Guin R. 2002. An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones. *Nucleic. Acids. Res.* **30:**2460–2468.

Caceres J.F., Kornblihtt A.R. 2002. Alternative splicing: Multiple control mechanisms and involvement in human disease. *Trends Genet*. **18:**186–193.

Cai J., Zhang J., Huang Y., Li Y. 2005. ATID: A web-oriented database for collection of publicly available alternative translational initiation events. *Bioinformatics* **21:**4312–4314.

Carninci P., Kasukawa T., Katayama S., Gough J., Frith M.C., Maeda N., Oyama R., Ravasi T., Lenhard B., Wells C., Kodzius R., Shimokawa K., Bajic V.B., Brenner S.E., Batalov S., Forrest A.R., Zavolan M., Davis M.J., Wilming L.G., Aidinis V., Allen J.E., Ambesi-Impiombato A., Apweiler R., Aturaliya R.N., Bailey T.L., Bansal M., Baxter L., Beisel K.W., Bersano T., Bono H., Chalk A.M., Chiu K.P., Choudhary V., Christoffels A., Clutterbuck D.R., Crowe M.L., Dalla E., Dalrymple B.P., de Bono B., Della Gatta G., di Bernardo D., Down T., Engstrom P., Fagiolini M., Faulkner G., Fletcher C.F., Fukushima T., Furuno M., Futaki S., Gariboldi M., Georgii-Hemming P., Gingeras T.R., Gojobori T., Green R.E., Gustincich S., Harbers M., Hayashi Y., Hensch T.K., Hirokawa N., Hill D., Huminiecki L., Iacono M., Ikeo K., Iwama A., Ishikawa T., Jakt M., Kanapin A., Katoh M., Kawasawa Y., Kelso J., Kitamura H., Kitano H., Kollias G., Krishnan S.P., Kruger A., Kummerfeld S.K., Kurochkin I.V., Lareau L.F., Lazarevic D., Lipovich L., Liu J., Liuni S., McWilliam S., Madan Babu M., Madera M., Marchionni L., Matsuda H., Matsuzawa S., Miki H., Mignone F., Miyake S., Morris K., Mottagui-Tabar S., Mulder N., Nakano N., Nakauchi H., Ng P., Nilsson R., Nishiguchi S., Nishikawa S., Nori F., Ohara O., Okazaki Y., Orlando V., Pang K.C., Pavan W.J., Pavesi G., Pesole G., Petrovsky N., Piazza S., Reed J., Reid J.F., Ring B.Z., Ringwald M., Rost B., Ruan Y., Salzberg S.L., Sandelin A., Schneider C., Schonbach C., Sekiguchi K., Semple C.A., Seno S., Sessa L., Sheng Y., Shibata Y., Shimada H., Shimada K., Silva D., Sinclair B., Sperling S., Stupka E., Sugiura K., Sultana R., Takenaka Y., Taki K., Tammoja K., Tan S.L., Tang S., Taylor M.S., Tegner J., Teichmann S.A., Ueda H.R., van Nimwegen E., Verardo R., Wei C.L., Yagi K., Yamanishi H., Zabarovsky E., Zhu S., Zimmer A., Hide W., Bult C., Grimmond S.M., Teasdale R.D., Liu E.T., Brusic V., Quackenbush J., Wahlestedt C., Mattick J.S., Hume D.A, Kai C., Sasaki D., Tomaru Y., Fukuda S., Kanamori-Katayama M., Suzuki M., Aoki J., Arakawa T., Iida J., Imamura K., Itoh M., Kato T., Kawaji H., Kawagashira N., Kawashima T., Kojima M., Kondo S., Konno H., Nakano K., Ninomiya N., Nishio T., Okada M., Plessy C., Shibata K., Shiraki T., Suzuki S., Tagami M., Waki K., Watahiki A., Okamura-Oho Y., Suzuki H., Kawai J., Hayashizaki Y. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:**1559–1563.

Carninci P., Sandelin A., Lenhard B., Katayama S., Shimokawa K., Ponjavic J., Semple C.A., Taylor M.S., Engstrom P.G., Frith M.C., Forrest A.R., Alkema W.B., Tan S.L., Plessy C., Kodzius R., Ravasi T., Kasukawa T., Fukuda S., Kanamori-Katayama M., Kitazume Y., Kawaji H., Kai C., Nakamura M., Konno H., Nakano K., Mottagui-Tabar S., Arner P., Chesi A., Gustincich S., Persichetti F., Suzuki H., Grimmond S.M., Wells C.A., Orlando V., Wahlestedt C., Liu E.T., Harbers M., Kawai J., Bajic V.B., Hume D.A., Hayashizaki, Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet*. **38:**626–635.

Cartegni L., Chew S.L., Krainer A.R. 2002. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet*. **3:**285–298.

Cartegni L., Krainer A.R. 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat. Genet*. **30:**377–384.

Cartegni L., Wang J., Zhu Z., Zhang M.Q., Krainer A.R. 2003. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic. Acids. Res.* **31:**3568–3571.

Castle J., Garrett-Engele P., Armour C.D., Duenwald S.J., Loerch P.M., Meyer M.R., Schadt E.E., Stoughton R., Parrish M.L., Shoemaker D.D., Johnson J.M. 2003. Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol*. **4:**R66.

Castrignano T., Rizzi R., Talamo I.G., De Meo P.D., Anselmo A., Bonizzoni, P., Pesole G. 2006. ASPIC: A web resource for alternative splicing prediction and transcript isoforms characterization. *Nucleic. Acids. Res.* **34:**W440–443.

Cawley S.L., Pachter L. 2003. HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics* **19(S2):**II36–II41.

Cheng J., Kapranov P., Drenkow J., Dike S., Brubaker S., Patel S., Long J., Stern D., Tammana H., Helt G., Sementchenko V., Piccolboni A., Bekiranov S., Bailey D.K., Ganesh M., Ghosh S., Bell I., Gerhard D.S., Gingeras T.R. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:**1149–1154.

Churbanov A., Pauley M., Quest D., Ali H. 2005. A method of precise mRNA/DNA homology-based gene structure prediction. *BMC Bioinformatics* **6:**261.

Clark T.A., Sugnet C.W., Ares M Jr. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296:**907–910.

Cline M.S., Blume J., Cawley S., Clark T.A., Hu J.S., Lu, G., Salomonis N., Wang H., Williams A. 2005. ANOSVA: A statistical method for detecting splice variation from expression data. *Bioinformatics* **21(S1):**i107–i115.

Cooper T.A. 2005. Use of minigene systems to dissect alternative splicing elements. *Methods* **37:**331–340.

Cuccurese M., Russo G., Russo A., Pietropaolo C. 2005. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic. Acids. Res.* **33:**5965–5977.

Cuperlovic-Culf M, Belacel N., Culf A.S., Ouellette R.J. 2006. Data analysis of alternative splicing microarrays. *Drug Discov. Today* **11:**983–990.

David A., Mabjeesh N., Azar I., Biton S., Engel S., Bernstein J., Romano J., Avidor Y., Waks T., Eshhar Z., Langer S.Z., Lifschitz-Mercer B., Matzkin H., Rotman G., Toporik A., Savitsky K., Mintz L. 2002. Unusual alternative splicing within the human kallikrein genes KLK2 and KLK3 gives rise to novel prostate-specific proteins. *J. Biol. Chem*. **277:**18084–18090.

Davis M.J., Hanson K.A., Clark F., Fink J.L., Zhang F., Kasukawa T., Kai C., Kawai J., Carninci P., Hayashizaki Y., Teasdale R.D. 2006. Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *PLoS Genet*. **2:**e46.

Dror G., Sorek R., Shamir R. 2005. Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics* **21:**897–901.

Fairbrother W.G., Yeo G.W., Yeh R., Goldstein P., Mawson M., Sharp P.A., Burge C.B. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic. Acids. Res.* **32:**W187–190.

Fan W., Khalid N., Hallahan A.R., Olson J.M., Zhao L.P. 2006. A statistical method for predicting splice variants between two groups of samples using GeneChip expression array data. *Theor. Biol. Med. Model*. **3:**19.

Faustino N.A., Cooper T.A. 2003. Pre-mRNA splicing and human disease. *Genes Dev*. **17:**419–437.

Fehlbaum P., Guihal C., Bracco L., Cochet O. 2005. A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic. Acids. Res.* **33:**e47.

Figeys D., Mc Broom L.D., Moran M.F. 2001. Mass spectrometry for the study of protein-protein interactions. *Methods* **24:**230–239.

Flicek, P., Brent M.R. 2006. Using several pair-wise informant sequences for *de novo* prediction of alternatively spliced transcripts. *Genome Biol*. **7(S1):**S8 1–9.

Florea L., Hartzell G., Zhang Z., Rubin G.M., Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:**967–974.

Forrest A.R., Taylor D.F., Crowe M.L., Chalk A.M., Waddell N.J., Kolle G., Faulkner G.J., Kodzius R., Katayama S., Wells C., Kai C., Kawai J., Carninci P., Hayashizaki Y., Grimmond S.M. 2006. Genome-wide review of transcriptional complexity in mouse protein kinases and phosphatases. *Genome Biol*. **7:**R5.

Frazer K.A., Pachter L., Poliakov A., Rubin E.M., Dubchak I. 2004. VISTA: Computational tools for comparative genomics. *Nucleic. Acids Res.* **32:**W273–279.

Garcia-Blanco M.A., Baraniak A.P., Lasda E.L. 2004. Alternative splicing in disease and therapy. *Nat. Biotechnol*. **22:**535–546.

Gerhard D.S., Wagner L., Feingold E.A., Shenmen C.M., Grouse L.H., Schuler G., Klein S.L., Old S., Rasooly R., Good P., Guyer M., Peck A.M., Derge J.G., Lipman D., Collins F.S., Jang W., Sherry S., Feolo M., Misquitta L., Lee E., Rotmistrovsky K., Greenhut S.F., Schaefer C.F., Buetow K., Bonner T.I., Haussler D., Kent J., Kiekhaus M., Furey T., Brent M., Prange C., Schreiber K., Shapiro N., Bhat N.K., Hopkins R.F., Hsie F., Driscoll T., Soares M.B., Casavant T.L., Scheetz T.E., Brownstein M.J., Usdin T.B., Toshiyuki S., Carninci P., Piao Y., Dudekula D.B., Ko M.S., Kawakami K.,

Suzuki Y., Sugano S., Gruber C.E., Smith M.R., Simmons B., Moore T., Waterman R., Johnson S.L., Ruan Y., Wei C.L., Mathavan S., Gunaratne P.H., Wu J., Garcia A.M., Hulyk S.W., Fuh E., Yuan Y., Sneed A., Kowis C., Hodgson A., Muzny D.M., McPherson J., Gibbs R.A., Fahey J., Helton E., Ketteman M., Madan A., Rodrigues S., Sanchez A., Whiting M., Madari A., Young A.C., Wetherby K.D., Granite S.J., Kwong P.N., Brinkley C.P., Pearson R.L., Bouffard G.G., Blakesly R.W., Green E.D., Dickson M.C., Rodriguez A.C., Grimwood J., Schmutz J., Myers R.M., Butterfield Y.S., Griffith M., Griffith O.L., Krzywinski M.I., Liao N., Morrin R., Palmquist D., Petrescu A.S., Skalska U., Smailus D.E., Stott J.M., Schnerch A., Schein J.E., Jones S.J., Holt R.A., Baross A., Marra M.A., Clifton S., Makowski K.A., Bosak S., Malek J. 2004. The status quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). *Genome Res.* **14:**2121–2127.

Goldstrohm A.C., Greenleaf A.L., Garcia-Blanco M.A. 2001. Co-transcriptional splicing of pre-messenger RNAs: Considerations for the mechanism of alternative splicing. *Gene.* **277:**31–47.

Goodson M.L., Jonas B.A., Privalsky M.L. 2005. Alternative mRNA splicing of SMRT creates functional diversity by generating corepressor isoforms with different affinities for different nuclear receptors. *J. Biol. Chem*. **280:**7493–7503.

Gowda M., Li H., Alessi J., Chen F., Pratt R., Wang G.L. 2006. Robust analysis of 5'-transcript ends (5'-RATE): A novel technique for transcriptome analysis and genome annotation. *Nucleic. Acids. Res.* **34:**e126.

Gupta S., Vingron M., Haas S.A. 2005. T-STAG: Resource and web-interface for tissue-specific transcripts and genes. *Nucleic. Acids. Res.* **33:**W654–658.

Gustincich S., Sandelin A., Plessy C., Katayama S., Simone R., Lazarevic D., Hayashizaki Y., Carninci P. 2006. The complexity of the mammalian transcripto.me. *J. Physiol*. **575:**321–332.

Harrison P.M., Kumar A., Lang N., Snyder M., Gerstein M. 2002. A question of size: The eukaryotic proteome and the problems in defining it. *Nucleic. Acids. Res.* **30:**1083–1090.

Hayashizaki Y., Carninci P. 2006. Genome Network and FANTOM3: Assessing the complexity of the transcriptome. *PLoS Genet*. **2:**e63.

Heber S., Alekseyev M., Sze S.H., Tang H., Pevzner P.A. 2002. Splicing graphs and EST assembly problem. *Bioinformatics* **18(S1):**S181–188.

Hicks M.J., Lam B.J., Hertel K.J. 2005. Analyzing mechanisms of alternative pre-mRNA splicing using *in vitro* splicing assays. *Methods* **37:**306–313.

Hiller M., Nikolajewa S., Huse K., Szafranski K., Rosenstiel P., Schuster S., Backofen R., Platzer M. 2006. TassDB: A database of alternative tandem splice sites. *Nucleic. Acids. Res.* **35:**188–192.

Hillier L.D., Lennon G., Becker M., Bonaldo M.F., Chiapelli B., Chissoe S., Dietrich N., DuBuque T., Favello A., Gish W., Hawkins M., Hultman M., Kucaba T., Lacy M., Le M., Le N., Mardis E., Moore B., Morris M., Parsons J., Prange C., Rifkin L., Rohlfing T., Schellenberg K., Marra M. 1996. Generation and analysis of 2,80,000 human expressed sequence tags. *Genome Res*. **6:**807–828.

Hillman R.T., Green R.E., Brenner S.E. 2004. An unappreciated role for RNA surveillance. *Genome. Biol*. **5:**R8.

Holste D., Huo G., Tung V., Burge C.B. 2006. HOLLYWOOD: A comparative relational database of alternative splicing. *Nucleic. Acids. Res.* **34:**D56–62.

Hu G.K., Madore S.J., Moldover B., Jatkoe T., Balaban D., Thomas J., Wang Y. 2001. Predicting splice variant from DNA chip expression data. *Genome Res.*. **11:**1237–1245.

Huang H.D., Horng J.T., Lee C.C., Liu B.J. 2003. ProSplicer: A database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. *Genome. Biol*. **4:**R29.

Huang H.D., Horng J.T., Lin F.M., Chang Y.C., Huang C.C. 2005. SpliceInfo: An information repository for mRNA alternative splicing in human genome. *Nucleic. Acids. Res.* **33:**D80–85.

Huang H.Y., Chien C.H., Jen K.H., Huang H.D. 2006. RegRNA: An integrated web server for identifying regulatory RNA motifs and elements. *Nucleic. Acids. Res.* **34:**W429–434.

Hubbard T., Andrews D., Caccamo M., Cameron G., Chen Y., Clamp M., Clarke L., Coates G., Cox T., Cunningham F., Curwen V., Cutts T., Down T., Durbin R., Fernandez-Suarez X.M., Gilbert J., Hammond M., Herrero J., Hotz H., Howe K., Iyer V., Jekosch K., Kahari A., Kasprzyk A., Keefe D., Keenan S., Kokocinsci F., London D., Longden I., McVicker G., Melsopp C., Meidl P., Potter S.,

Proctor G., Rae M., Rios D., Schuster M., Searle S., Severin J., Slater G., Smedley D., Smith J., Spooner W., Stabenau A., Stalker J., Storey R., Trevanion S., Ureta-Vidal A., Vogel J., White S., Woodwark C.,Birney E. 2005. Ensembl 2005. *Nucleic. Acids. Res.* **33:**D447–453.

Hughes T.A. 2006. Regulation of gene expression by alternative untranslated regions. *Trends Genet*. **22:**119–122.

Hui L., Zhang X., Wu X., Lin Z., Wang Q., Li Y., Hu G. 2004. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* **23:**3013–3023.

Imanishi T., Itoh T., Suzuki Y., O'Donovan C., Fukuchi S., Koyanagi K.O., Barrero R.A., Tamura T., Yamaguchi-Kabata Y., Tanino M., Yura K., Miyazaki S., Ikeo K., Homma K., Kasprzyk A., Nishikawa T., Hirakawa M., Thierry-Mieg J., Thierry-Mieg D., Ashurst J., Jia L., Nakao M., Thomas M.A., Mulder N., Karavidopoulou Y., Jin L., Kim S., Yasuda T., Lenhard B., Eveno E., Suzuki Y., Yamasaki C., Takeda J., Gough C., Hilton P., Fujii Y., Sakai H., Tanaka S., Amid C., Bellgard M., Bonaldo Mde F., Bono H., Bromberg S.K., Brookes A.J., Bruford E., Carninci P., Chelala C., Couillault C., de Souza S.J., Debily M.A., Devignes M.D., Dubchak I., Endo T., Estreicher A., Eyras E., Fukami-Kobayashi K., Gopinath G.R., Graudens E., Hahn Y., Han M., Han Z.G., Hanada K., Hanaoka H., Harada E., Hashimoto K., Hinz U., Hirai M., Hishiki T., Hopkinson I., Imbeaud S., Inoko H., Kanapin A., Kaneko Y., Kasukawa T., Kelso J., Kersey P., Kikuno R., Kimura K., Korn B., Kuryshev V., Makalowska I., Makino T., Mano S., Mariage-Samson R., Mashima J., Matsuda H., Mewes H.W., Minoshima S., Nagai K., Nagasaki H., Nagata N., Nigam R., Ogasawara O., Ohara O., Ohtsubo M., Okada N., Okido T., Oota S., Ota M., Ota T., Otsuki T., Piatier-Tonneau D., Poustka A., Ren S.X., Saitou N., Sakai K., Sakamoto S., Sakate R., Schupp I., Servant F., Sherry S., Shiba R., Shimizu N., Shimoyama M., Simpson A.J., Soares B., Steward C., Suwa M., Suzuki M., Takahashi A., Tamiya G., Tanaka H., Taylor T., Terwilliger J.D., Unneberg P., Veeramachaneni V., Watanabe S., Wilming L., Yasuda N., Yoo H.S., Stodolsky M., Makalowski W., Go M., Nakai K., Takagi T., Kanehisa M., Sakaki Y., Quackenbush J., Okazaki Y., Hayashizaki Y., Hide W., Chakraborty R., Nishikawa K., Sugawara H., Tateno Y., Chen Z., Oishi M., Tonellato P., Apweiler R., Okubo K., Wagner L., Wiemann S., Strausberg R.L., Isogai T., Auffray C., Nomura N., Gojobori T., Sugano S. 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*. **2:**e162.

Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D, Antonellis K.J., Scherf U., Speed T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:**249–264.

Johnson J.M., Castle J., Garrett-Engele P, Kan Z., Loerch P.M., Armour C.D., Santos R., Schadt E.E., Stoughton R., Shoemaker D.D. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:**2141–2144.

Johnson J.M., Edwards S., Shoemaker D., Schadt E.E. 2005. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet*. **21:**93–102.

Jones S.J. 2005. Prediction of genomic functional elements. *Annu. Rev. Genomics Hum. Genet*. **7:**315–338.

Kalnina Z., Zayakin P., Silina K., Line A. 2005. Alterations of pre-mRNA splicing in cancer. *Genes, Chromosomes & Cancer* **42:**342–357.

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

Kan Z., Rouchka E.C., Gish W.R., States D.J. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res*. **11:**889–900.

Kapranov P., Cawley S.E., Drenkow J., Bekiranov S., Strausberg R.L. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:**916–919.

Kemp P.R., Ellis P.D., Smith C.W. 2005. Visualization of alternative splicing *in vivo*. *Methods* **37:**360–367.

Kent W.J. 2002. BLAT- the BLAST-like alignment tool. *Genome Res*. **12:**656–664.

Kim N., Alekseyenko A.V., Roy M., Lee C. 2006. The ASAP II database: Analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic. Acids. Res.* **35:**D93–98.

Kim N., Lim D., Lee S., Kim H. 2005. ASePCR: Alternative splicing electronic RT-PCR in multiple tissues and organs. *Nucleic. Acids. Res.* **33:**W681–685.

Kim N., Shin S., Lee S. 2005. ECgene: Genome-based EST clustering and gene modeling for alternative splicing. *Genome Res*. **15:**566–576.

Kim P., Kim N., Lee Y., Kim B., Shin Y., Lee S. 2005. ECgene: Genome annotation for alternative splicing. *Nucleic. Acids. Res.* **33:**D75–79.

Kirschbaum-Slager N, Lopes G.M., Galante P.A., Riggins G.J., de Souza S.J. 2004. Splicing factors are differentially expressed in tumors. *Genet. Mol. Res*. **3:**512–520.

Kirschbaum-Slager N, Parmigiani R.B., Camargo A.A., de Souza S.J. 2005. Identification of human exons over-expressed in tumors through the use of genome and expressed sequence data. *Physiol. Genomics* **21:**423–432.

Kopelman N.M., Lancet D., Yanai I. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat. Genet*. **37:**588–589.

Kornblihtt A.R. 2005. Promoter usage and alternative splicing. *Curr. Opin. Cell Biol*. **17:**262–268.

Kostadinov R., Malhotra N., Viotti M., Shine R., D'Antonio L., Bagga P. 2006. GRSDB: A database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. *Nucleic. Acids. Res*. **34:**D119–124.

Kuhn R.M., Karolchik D., Zweig A.S., Trumbower H., Thomas D.J., Thakkapallayil A., Sugnet C.W., Stanke M., Smith K.E., Siepel A., Rosenbloom K.R., Rhead B., Raney B.J., Pohl A., Pedersen J.S., Hsu F., Hinrichs A.S., Harte R.A., Diekhans M., Clawson H., Bejerano G., Barber G.P., Baertsch R., Haussler D., Kent W.J. 2006. The UCSC genome browser database: Update 2007. *Nucleic. Acids. Res.* (in press).

Kuo B.Y., Chen Y., Bohacec S., Johansson O., Wasserman W.W., Simpson E.M. 2006. SAGE2Splice: Unmapped SAGE tags reveal novel splice junctions. *PLoS Comput. Biol*. **2:**e34.

Kuroyanagi H., Kobayashi T., Mitani S., Hagiwara M. 2006. Transgenic alternative-splicing reporters reveal tissue-specific expression profiles and regulation mechanisms *in vivo*. *Nat. Methods* **3:**909–915.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J.C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R.H., Wilson R.K., Hillier L.W., McPherson J.D., Marra M.A., Mardis E.R., Fulton L.A., Chinwalla A.T., Pepin K.H., Gish W.R., Chissoe S.L., Wendl M.C., Delehaunty K.D., Miner T.L., Delehaunty A., Kramer J.B., Cook L.L., Fulton R.S., Johnson D.L., Minx P.J., Clifton S.W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J.F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., Gibbs R.A., Muzny D.M., Scherer S.E., Bouck J.B., Sodergren E.J., Worley K.C., Rives C.M., Gorrell J.H., Metzker M.L., Naylor S.L., Kucherlapati R.S., Nelson D.L., Weinstock G.M., Sakaki Y., Fujiyama A., Hattori M., Yada T., Toyoda A., Itoh T., Kawagoe C., Watanabe H., Totoki Y., Taylor T., Weissenbach J., Heilig R., Saurin W., Artiguenave F., Brottier P., Bruls T., Pelletier E., Robert C., Wincker P., Smith D.R., Doucette-Stamm L., Rubenfield M., Weinstock K., Lee H.M., Dubois J., Rosenthal A., Platzer M., Nyakatura G., Taudien S., Rump A., Yang H., Yu J., Wang J., Huang G., Gu J., Hood L., Rowen L., Madan A., Qin S., Davis R.W., Federspiel N.A., Abola A.P., Proctor M.J., Myers R.M., Schmutz J., Dickson M., Grimwood J., Cox D.R., Olson M.V., Kaul R., Raymond C., Shimizu N., Kawasaki K., Minoshima S., Evans G.A., Athanasiou M., Schultz R., Roe B.A., Chen F., Pan H., Ramser J., Lehrach H., Reinhardt R., McCombie W.R., de la Bastide M., Dedhia N., Blocker H., Hornischer K., Nordsiek G., Agarwala R., Aravind L., Bailey J.A., Bateman A., Batzoglou S., Birney E., Bork P., Brown D.G., Burge C.B., Cerutti L., Chen H.C., Church D., Clamp M., Copley R.R., Doerks T., Eddy S.R., Eichler E.E., Furey T.S., Galagan J., Gilbert J.G., Harmon C., Hayashizaki Y., Haussler D., Hermjakob H., Hokamp K., Jang W., Johnson L.S., Jones T.A., Kasif S., Kaspryzk A., Kennedy S., Kent W.J., Kitts P., Koonin E.V., Korf I., Kulp D., Lancet D., Lowe T.M., McLysaght A., Mikkelsen T., Moran J.V., Mulder N., Pollara V.J., Ponting C.P., Schuler G., Schultz J., Slater G., Smit A.F., Stupka E., Szustakowski J., Thierry-Mieg D., Thierry-

Mieg J., Wagner L., Wallis J., Wheeler R., Williams A., Wolf Y.I., Wolfe K.H., Yang S.P., Yeh R.F., Collins F., Guyer M.S., Peterson J., Felsenfeld A., Wetterstrand K.A., Patrinos A., Morgan M.J., de Jong P., Catanese J.J., Osoegawa K., Shizuya H., Choi S., Chen Y.J. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:**860–921.

Lareau L.F., Green R.E., Bhatnagar R.S., Brenner S.E. 2004. The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol*. **14:**273–282.

Le K., Mitsouras K., Roy M., Wang Q., Xu Q., Nelson S.F., Lee C. 2004. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic. Acids. Res.* **32:**e180.

Le Texier V., Riethoven J.J., Kumanduri V., Gopalakrishnan C., Lopez F., Gautheret D., Thanaraj T.A. 2006. AltTrans: Transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics* **7:**169.

Leamon J.H., Braverman M.S., Rothberg J.M. 2007. High-throughput, massively parallel DNA sequencing technology for the era of personalized medicine. *Gene Therapy and Regulation* **3:**15–31.

Lee C., Roy M. 2004. Analysis of alternative splicing with microarrays: Successes and challenges. *Genome Biol*. **5:**231.

Lewis B.P., Green R.E., Brenner S.E. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. USA* **100:**189–192.

Li C., Kato M., Shiue L., Shively J.E., Ares M. Jr., Lin R.J. 2006. Cell type and culture condition-dependent alternative splicing in human breast cancer cells revealed by splicing-sensitive microarrays. *Cancer Res*. **66:**1990–1999.

Licatalosi D.D., Darnell R.B. 2006. Splicing regulation in neurologic disease. *Neuron.* **52:**93–101.

Lu C., Tej S.S., Luo S., Haudenschild C.D., Meyers B.C., Green P.J. 2005. Elucidation of the small RNA component of the transcriptome. *Science* **309:**1567–1569.

Lu J., Lal A., Merriman B., Nelson S., Riggins G. 2004. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics* **84:**631–636.

Lyddy J. 2002. ExonHit Therapeutics. *Pharmacogenomics* **3:**843–846.

Major S.M., Nishizuka S., Morita D., Rowland R., Sunshine M., Shankavaram U., Washburn F., Asin D., Kouros-Mehr H., Kane D., Weinstein J.N. 2006. AbMiner: A bioinformatic resource on available monoclonal antibodies and corresponding gene identifiers for genomic, proteomic, and immunologic studies. *BMC Bioinformatics* **7:**192.

Mangasarian A. 2005. Alternative RNA splicing and drug target identification. *IDrugs* **8:**725–729.

Maniatis T., Reed R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416:**499–506.

Maniatis T., Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:**236–243.

Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L., Jarvie T.P., Jirage K.B., Kim J.B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F., Rothberg J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:**376–380.

Matlin A.J., Clark F., Smith C.W. 2005. Understanding alternative splicing: Towards a cellular code. *Nat. Rev. Mol. Cell Biol*. **6:**386–398.

Metzker M.L. 2005. Emerging technologies in DNA sequencing. *Genome Res*. **15:**1767–1776.

Mironov A.A., Fickett J.W., Gelfand M.S. 1999. Frequent alternative splicing of human genes. *Genome Res*. **9:**1288–1293.

Modrek B., Lee C.J. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet*. **34:**177–180.

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

Nagasaki H., Arita M., Nishizawa T., Suwa M., Gotoh O. 2006. Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics* **22:**1211–1216.

Nanjundan M., Zhang F., Schmandt R., Smith-McCune K., Mills G.B. 2006. Identification of a novel splice variant of AML1b in ovarian cancer patients conferring loss of wild-type tumor suppressive functions. *Oncogene* (in press).

Ng P., Tan J.J., Ooi H.S., Lee Y.L., Chiu K.P., Fullwood M.J., Srinivasan K.G., Perbost C., Du L., Sung W.K., Wei C.L., Ruan Y. 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:**e84.

Ng P., Wei C.L., Sung W.K., Chiu K.P., Lipovich L., Ang C.C., Gupta S., Shahab A., Ridwan A., Wong C.H., Liu E.T., Ruan Y. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* **2:**105–111.

Nielsen K.L., Hogh A.L., Emmersen J. 2006. DeepSAGE — digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic. Acids. Res.* **34:**e133.

Noh S.J., Lee K., Paik H., Hur C.G. 2006. TISA: Tissue-specific alternative splicing in human and mouse genes. *DNA Res*. **13:**229–243.

Nuwaysir E.F., Huang W., Albert T.J., Singh J., Nuwaysir K., Pitas, A., Richmond, T., Gorski, T., Berg, J.P., Ballin, J., McCormick, M., Norton, J., Pollock, T., Sumwalt, T., Butcher, L., Porter, D., Molla, M., Hall, C., Blattner, F., Sussman, M.R., Wallace, R.L., Cerrina, F. and Green, R.D. 2002. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res*. **12:**1749–1755.

Ohler U., Shomron N., Burge C.B. 2005. Recognition of unknown conserved alternatively spliced exons. *PLoS Comput. Biol*. **1:**113–122.

Okazaki Y., Furuno M., Kasukawa T., Adachi J., Bono H., Kondo S., Nikaido I., Osato N., Saito R., Suzuki H., Yamanaka I., Kiyosawa H., Yagi K., Tomaru Y., Hasegawa Y., Nogami A., Schonbach C., Gojobori T., Baldarelli R., Hill D.P., Bult C., Hume D.A., Quackenbush J., Schriml L.M., Kanapin A., Matsuda H., Batalov S., Beisel K.W., Blake J.A., Bradt D., Brusic V., Chothia C., Corbani L.E., Cousins S., Dalla E., Dragani T.A., Fletcher C.F., Forrest A., Frazer K.S., Gaasterland T., Gariboldi M., Gissi C., Godzik A., Gough J., Grimmond S., Gustincich S., Hirokawa N., Jackson I.J., Jarvis E.D., Kanai A., Kawaji H., Kawasawa Y., Kedzierski R.M., King B.L., Konagaya A., Kurochkin I.V., Lee Y., Lenhard B., Lyons P.A., Maglott D.R., Maltais L., Marchionni L., McKenzie L., Miki H., Nagashima T., Numata K., Okido T., Pavan W.J., Pertea G., Pesole G., Petrovsky N., Pillai R., Pontius J.U., Qi D., Ramachandran S., Ravasi T., Reed J.C., Reed D.J., Reid J., Ring B.Z., Ringwald M., Sandelin A., Schneider C., Semple C.A., Setou M., Shimada K., Sultana R., Takenaka Y., Taylor M.S., Teasdale R.D., Tomita M., Verardo R., Wagner L., Wahlestedt C., Wang Y., Watanabe Y., Wells C., Wilming L.G., Wynshaw-Boris A., Yanagisawa M., Yang I., Yang L., Yuan Z., Zavolan M., Zhu Y., Zimmer A., Carninci P., Hayatsu N., Hirozane-Kishikawa T., Konno H., Nakamura M., Sakazume N., Sato K., Shiraki T., Waki K., Kawai J., Aizawa K., Arakawa T., Fukuda S., Hara A., Hashizume W., Imotani K., Ishii Y., Itoh M., Kagawa I., Miyazaki A., Sakai K., Sasaki D., Shibata K., Shinagawa A., Yasunishi A., Yoshino M., Waterston R., Lander E.S., Rogers J., Birney E., Hayashizaki Y. 2002. Analysis of the mouse transcriptome based on functional annotation of 60770 full-length cDNAs. *Nature* **420:**563–573.

Ota T., Suzuki Y., Nishikawa T., Otsuki T., Sugiyama T., Irie R., Wakamatsu A., Hayashi K., Sato H., Nagai K., Kimura K., Makita H., Sekine M., Obayashi M., Nishi T., Shibahara T., Tanaka T., Ishii S., Yamamoto J., Saito K., Kawai Y., Isono Y., Nakamura Y., Nagahari K., Murakami K., Yasuda T., Iwayanagi T., Wagatsuma M., Shiratori A., Sudo H., Hosoiri T., Kaku Y., Kodaira H., Kondo H., Sugawara M., Takahashi M., Kanda K., Yokoi T., Furuya T., Kikkawa E., Omura Y., Abe K., Kamihara K., Katsuta N., Sato K., Tanikawa M., Yamazaki M., Ninomiya K., Ishibashi T., Yamashita H., Murakawa K., Fujimori K., Tanai H., Kimata M., Watanabe M., Hiraoka S., Chiba Y., Ishida S., Ono Y., Takiguchi S., Watanabe S., Yosida M., Hotuta T., Kusano J., Kanehori K., Takahashi-Fujii A., Hara H., Tanase T.O., Nomura Y., Togiya S., Komai F., Hara R., Takeuchi K., Arita M., Imose N., Musashino K., Yuuki H., Oshima A., Sasaki N., Aotsuka S., Yoshikawa Y., Matsunawa H., Ichihara T., Shiohata N., Sano S., Moriya S., Momiyama H., Satoh N., Takami S., Terashima Y., Suzuki O., Nakagawa S., Senoh A., Mizoguchi H., Goto Y., Shimizu F., Wakebe H., Hishigaki H., Watanabe T., Sugiyama A., Takemoto M., Kawakami B., Yamazaki M., Watanabe K., Kumagai A., Itakura S., Fukuzumi Y., Fujimori Y., Komiyama M., Tashiro H., Tanigami A., Fujiwara T., Ono T., Yamada K., Fujii Y., Ozaki K., Hirao M., Ohmori Y., Kawabata A., Hikiji T., Kobatake N., Inagaki H., Ikema Y., Okamoto S., Okitani R., Kawakami T., Noguchi S., Itoh T., Shigeta K., Senba T., Matsumura K., Nakajima Y., Mizuno T., Morinaga M., Sasaki M., Togashi T., Oyama M., Hata

H., Watanabe M., Komatsu T., Mizushima-Sugano J., Satoh T., Shirai Y., Takahashi Y., Nakagawa K., Okumura K., Nagase T., Nomura N., Kikuchi H., Masuho Y., Yamashita R., Nakai K., Yada T., Nakamura Y., Ohara O., Isogai T., Sugano S. 2004. Complete sequencing and characterization of 21243 full-length human cDNAs. *Nat. Genet*. **36:**40–45.

Paddison P.J., Silva J.M. Conklin D.S., Schlabach M. Li M., Aruleba S. Balija V., O'Shaughnessy A. Gnoj L., Scobie K., Chang K. Westbrook T., Cleary M. Sachidanandam R., McCombie W.R. Elledge S.J., Hannon G.J. 2004. A resource for large-scale RNA-interference-based screens in mammals. *Nature* **428:**427–431.

Pagani F. Baralle F.E. 2004. Genomic variants in exons and introns: Identifying the splicing spoilers. *Nat. Rev. Genet*. **5:**389–396.

Pan Q., Bakowski M.A., Morris Q., Zhang W., Frey B.J., Hughes T.R., Blencowe B.J. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet*. **21:**73–77.

Pan Q., Saltzman A.L., Kim Y.K., Misquitta C., Shai O., Maquat L.E., Frey B.J., Blencowe B.J. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev*. **20:**153–158.

Pan Q., Shai O., Misquitta C., Zhang W., Saltzman A.L., Mohammad N., Babak T., Siu H., Hughes T.R., Morris Q.D., Frey B.J., Blencowe B.J. 2004. Revealing global regulatory features of Mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell* **16:**929–941.

Philipps D.L., Park J.W., Graveley B.R. 2004. A computational and experimental approach toward *a priori* identification of alternatively spliced exons. *RNA* **10:**1838–1844.

Ravasi T., Huber T., Zavolan M., Forrest A., Gaasterland T., Grimmond S., Hume D.A. 2003. Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. *Genome Res*. **13:**1430–1442.

Redkar R., Burzio L., Haines D., Conzone S. 2006. Microarray technology: Past, present and future. In Thangadurai D., Tang W. and Pullaiah T. (eds.), *Genes, Genomes and Genomics*, Vol. 1., Regency Publications, New Delhi, pp. 1–39.

Relogio A., Ben-Dov C, Baum M., Ruggiu M., Gemund C., Benes V., Darnell R.B., Valcarcel J. 2004. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J. Biol. Chem*. **280:**4779–4784.

Resch A., Xing Y., Modrek B., Gorlick M., Riley R., Lee C. 2004. Assessing the impact of alternative splicing on domain interactions in the human proteome. *J. Proteome. Res.* **3:**76–83.

Roberts G.C., Smith C.W. 2002. Alternative splicing: Combinatorial output from the genome. *Curr. Opin. Chem. Biol*. **6:**375–383.

Ruan Y., Le Ber P., Ng H.H., Liu E.T. 2004. Interrogating the transcriptome. *Trends Biotechnol*. **22:**23–30.

Saha S., Sparks A.B., Rago C., Akmaev V., Wang C.J. 2002. Using the transcriptome to annotate the genome. *Nat. Biotechnol*. **20:**508–512.

Salesse S., Dylla S.J., Verfaillie C.M. 2004. p210BCR/ABL-induced alteration of pre-mRNA splicing in primary human CD34+ hematopoietic progenitor cells. *Leukemia* **18:**727–733.

Schadt E.E., Edwards S.W., GuhaThakurta D., Holder D., Ying L., Svetnik V., Leonardson A., Hart K.W., Russell A., Li G., Cavet G., Castle J., McDonagh P., Kan Z., Chen R., Kasarskis A., Margarint M., Caceres R.M., Johnson J.M., Armour C.D., Garrett-Engele P.W., Tsinoremas N.F., Shoemaker D.D. 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol*. **5:**R73.

Schmitt A.O., Specht T., Beckmann G., Dahl E., Pilarsky C.P., Hinzmann B., Rosenthal A. 1999. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic. Acids. Res.* **27:**4251–4260.

Schmucker D., Clemens J.C., Shu H., Worby C.A., Xiao J., Muda M., Dixon J.E., Zipursky S.L. 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101:**671–684.

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

Shah P.K., Jensen L.J., Boue S., Bork P. 2005. Extraction of transcript diversity from scientific literature. *PLoS Comput. Biol*. **1:**e10.

Shai O., Morris Q.D., Blencowe B.J., Frey B.J. 2006. Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics* **22:**606–613.

Sharp P.A. 1994. Split genes and RNA splicing. *Cell* **77:**805–815.

Shiraki T., Kondo S., Katayama S., Waki K., Kasukawa T., Kawaji H., Kodzius R., Watahiki A., Nakamura M., Arakawa T., Fukuda S., Sasaki D., Podhajska A., Harbers M., Kawai J., Carninci P., Hayashizaki Y. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100:**15776–15781.

Siddiqui A.S., Delaney A.D., Schnerch A., Griffith O.L., Jones S.J., Marra M.A. 2006. Sequence biases in large scale gene expression profiling data. *Nucleic. Acids. Res.* **34:**e83.

Siddiqui A.S., Khattra J., Delaney A.D., Zhao Y., Astell C., Asano J., Babakaiff R., Barber S., Beland J., Bohacec S., Brown-John M., Chand S., Charest D., Charters A.M., Cullum R., Dhalla N., Featherstone R., Gerhard D.S., Hoffman B., Holt R.A., Hou J., Kuo B.Y., Lee L.L., Lee S., Leung D., Ma K., Matsuo C., Mayo M., McDonald H., Prabhu A.L., Pandoh P., Riggins G.J., de Algara T.R., Rupert J.L., Smailus D., Stott J., Tsai M., Varhol R., Vrljicak P., Wong D., Wu M.K., Xie Y.Y., Yang G., Zhang I., Hirst M., Jones S.J., Helgason C.D., Simpson E.M., Hoodless P.A., Marra M.A. 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. USA* **102:**18485–18490.

Smith P.J., Zhang C., Wang J., Chew S.L., Zhang M.Q., Krainer A.R. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum. Mol. Genet*. **15:**2490–2508.

Soller M. 2006. Pre-messenger RNA processing and its regulation: A genomic perspective. *Cell Mol. Life Sci*. **63:**796–819.

Sorek R., Ast G. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res*. **13:**1631–1637.

Sorek R., Shemesh R., Cohen Y., Basechess O., Ast G., Shamir R. 2004. A non-EST-based method for exon-skipping prediction. *Genome Res*. **14:**1617–1623.

Srebrow A., Kornblihtt A.R. 2006. The connection between splicing and cancer. *J. Cell Sci*. **119:**2635–2641.

Srinivasan K., Shiue L., Hayes J.D., Centers R., Fitzwater S., Loewen R., Edmondson L.R., Bryant J., Smith M., Rommelfanger C., Welch V., Clark T.A., Sugnet C.W., Howe K.J., Mandel-Gutfreund Y., Ares M. Jr. 2005. Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods* **37:**345–359.

Stamm S., Ben-Ari S., Rafalska I., Tang Y., Zhang Z. 2005. Function of alternative splicing. *Gene.* **344:**1–20.

Stamm S., Riethoven J.J., Le Texier V., Gopalakrishnan C., Kumanduri V., Tang Y., Barbosa-Morais N.L., Thanaraj T.A. 2006. ASD: A bioinformatics resource on alternative splicing. *Nucleic. Acids. Res.* **34:**D46–55.

Stanke M., Keller O., Gunduz I., Hayes A., Waack S., Morgenstern B. 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic. Acids. Res.* **34:**W435–439.

Stenson P.D., Ball E.V., Mort M., Phillips A.D., Shiel J.A., Thomas N.S., Abeysinghe S., Krawczak M., Cooper D.N. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat*. **21:**577–581.

Stoilov P., Meshorer E., Gencheva M., Glick D., Soreq H., Stamm S. 2002. Defects in pre-mRNA processing as causes of and predisposition to diseases. *DNA Cell Biol*. **21:**803–818.

Stolc V., Gauhar Z., Mason C., Halasz G., van Batenburg M.F., Rifkin S.A., Hua S., Herreman T., Tongprasit W., Barbano P.E., Bussemaker H.J., White K.P. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306:**655–660.

Strausberg R.L. 2001. The Cancer Genome Anatomy Project: New resources for reading the molecular signatures of cancer. *J. Pathol*. **195:**31–40.

Strausberg R.L., Feingold E.A., Klausner R.D., Collins F.S. 1999. The mammalian gene collection. *Science* **286:**455–457.

Su Z., Wang J., Yu J., Huang X., Gu X. 2006. Evolution of alternative splicing after gene duplication. *Genome Res*. **16:**182–189.

Sugnet C.W., Srinivasan K., Clark T.A., O'Brien G., Cline M.S., Wang H., Williams A., Kulp D., Blume J.E., Haussler D., Ares M. Jr. 2006. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput. Biol*. **2:**e4.

Suzuki Y., Yamashita R., Sugano S., Nakai K. 2004. DBTSS, DataBase of Transcriptional Start Sites: Progress report 2004. *Nucleic. Acids. Res.* **32:**D78–81.

Takeda J., Suzuki Y., Nakao M., Barrero R.A., Koyanagi K.O., Jin L., Motono C., Hata H., Isogai T., Nagai K., Otsuki T., Kuryshev V., Shionyu M., Yura K., Go M., Thierry-Mieg J., Thierry-Mieg D., Wiemann S., Nomura N., Sugano S., Gojobori T., Imanishi T. 2006. Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic. Acids. Res.* **34:**3917–3928.

Teraoka S.N., Telatar M., Becker-Catania S, Liang T., Onengut S., Tolun A., Chessa L., Sanal O., Bernatowska E., Gatti R.A., Concannon P. 1999. Splicing defects in the ataxia-telangiectasia gene, ATM: Underlying mutations and consequences. *Am. J. Hum. Genet*. **64:**1617–1631.

Thanaraj T.A., Clark F., Muilu J. 2003. Conservation of human alternative splice events in mouse. *Nucleic. Acids. Res.* **31:**2544–2552.

Thill G., Castelli V., Pallud S., Salanoubat M., Wincker P., de la Grange P., Auboeuf D., Schachter V., Weissenbach J. 2006. ASEtrap: A biological method for speeding up the exploration of spliceomes. *Genome Res*. **16:**776–786.

Thomas R.K., Nickerson E., Simons J.F., Janne P.A., Tengs T., Yuza Y., Garraway L.A., LaFramboise T., Lee J.C., Shah K., O'Neill K., Sasaki H., Lindeman N., Wong K.K., Borras A.M., Gutmann E.J., Dragnev K.H., DeBiasi R., Chen T.H., Glatt K.A., Greulich H., Desany B., Lubeski C.K., Brockman W., Alvarez P., Hutchison S.K., Leamon J.H., Ronan M.T., Turenchalk G.S., Egholm M., Sellers W.R., Rothberg J.M., Meyerson M. 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat. Med*. **12:**852–855.

Ule J., Jensen K., Mele A., Darnell R.B. 2005. CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods* **37:**376–386.

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

Ule J., Ule A., Spencer J., Williams A., Hu J.S., Cline M., Wang H., Clark T., Fraser C., Ruggiu M., Zeeberg B.R., Kane D., Weinstein J.N., Blume J., Darnell R.B. 2005. Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet*. **37:**844–852.

Usuka J., Zhu W., Brendel V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16:**203–211.

van Nimwegen E., Paul N., Sheridan R., Zavolan M. 2006. SPA: A probabilistic algorithm for spliced alignment. *PLoS Genet*. **2:**e24.

van Ruissen F., Ruijter J.M., Schaaf G.J., Asgharnegad L., Zwijnenburg D.A., Kool M., Baas F. 2005. Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics* **6:**91.

Vanhoutteghem A., Djian P. 2006. The human basonuclin 2 gene has the potential to generate nearly 90,000 mRNA isoforms encoding over 2000 different proteins. *Genomics* **89:**44–58.

Vegran F., Boidot R., Oudin C., Riedinger J.M., Bonnetain F., Lizard-Nacol S. 2006. Overexpression of caspase-3s splice variant in locally advanced breast carcinoma is associated with poor response to neoadjuvant chemotherapy. *Clin. Cancer Res*. **12:**5794–5800.

Velculescu V.E., Zhang L., Vogelstein B., Kinzler K.W. 1995. Serial analysis of gene expression. *Science* **270:**484–487.

Venables J.P., Burn J. 2006. EASI — enrichment of alternatively spliced isoforms. *Nucleic. Acids. Res.* **34:**e103.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., Nelson C., Broder S., Clark A.G., Nadeau J., McKusick V.A., Zinder N., Levine A.J., Roberts R.J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A.E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T.J., Higgins M.E., Ji R.R., Ke Z., Ketchum K.A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G.V., Milshina N., Moore H.M., Naik A.K., Narayan V.A., Neelam B., Nusskern D., Rusch D.B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R.,

Xiao C., Yan C., Yao A., Ye J., Zhan M., Zhang W., Zhang H., Zhao Q., Zheng L., Zhong F., Zhong W., Zhu S., Zhao S., Gilbert D., Baumhueter S., Spier G., Carter C., Cravchik A., Woodage T., Ali F., An H., Awe A., Baldwin D., Baden H., Barnstead M., Barrow I., Beeson K., Busam D., Carver A., Center A., Cheng M.L., Curry L., Danaher S., Davenport L., Desilets R., Dietz S., Dodson K., Doup L., Ferriera S., Garg N., Gluecksmann A., Hart B., Haynes J., Haynes C., Heiner C., Hladun S., Hostin D., Houck J., Howland T., Ibegwam C., Johnson J., Kalush F., Kline L., Koduru S., Love A., Mann F., May D., McCawley S., McIntosh T., McMullen I., Moy M., Moy L., Murphy B., Nelson K., Pfannkoch C., Pratts E., Puri V., Qureshi H., Reardon M., Rodriguez R., Rogers Y.H., Romblad D., Ruhfel B., Scott R., Sitter C., Smallwood M., Stewart E., Strong R., Suh E., Thomas R., Tint N.N., Tse S., Vech C., Wang G., Wetter J., Williams S., Williams M., Windsor S., Winn-Deen E., Wolfe K., Zaveri J., Zaveri K., Abril J.F., Guigo R., Campbell M.J., Sjolander K.V., Karlak B., Kejariwal A., Mi H., Lazareva B., Hatton T., Narechania A., Diemer K., Muruganujan A., Guo N., Sato S., Bafna V., Istrail S., Lippert R., Schwartz R., Walenz B., Yooseph S., Allen D., Basu A., Baxendale J., Blick L., Caminha M., Carnes-Stine J., Caulk P., Chiang Y.H., Coyne M., Dahlke C., Mays A., Dombroski M., Donnelly M., Ely D., Esparham S., Fosler C., Gire H., Glanowski S., Glasser K., Glodek A., Gorokhov M., Graham K., Gropman B., Harris M., Heil J., Henderson S., Hoover J., Jennings D., Jordan C., Jordan J., Kasha J., Kagan L., Kraft C., Levitsky A., Lewis M., Liu X., Lopez J., Ma D., Majoros W., McDaniel J., Murphy S., Newman M., Nguyen T., Nguyen N., Nodell M., Pan S., Peck J., Peterson M., Rowe W., Sanders R., Scott J., Simpson M., Smith T., Sprague A., Stockwell T., Turner R., Venter E., Wang M., Wen M., Wu D., Wu M., Xia A., Zandieh A., Zhu X. 2001. The sequence of the human genome. *Science* **291:**1304–1351.

Wang H., Hubbell E., Hu J.S., Mei G., Cline M., Lu G., Clark T., Siani-Rose M.A., Ares M., Kulp D.C., Haussler D. 2003. Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics* **19(S1):**i315–322.

Wang P., Yan B., Guo J.T., Hicks C., Xu Y. 2005. Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proc. Natl. Acad. Sci. USA* **102:**18920–18925.

Wang Z., Rolish M.E., Yeo G., Tung V., Mawson M., Burge C.B. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119:**831–845.

Watahiki A., Waki K., Hayatsu N., Shiraki T., Kondo S., Nakamura M., Sasaki D., Arakawa T., Kawai J., Harbers M., Hayashizaki Y., Carninci P. 2004. Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *Nat. Methods* **1:**233–239.

Watson F.L., Puttmann-Holgado R, Thomas F., Lamar D.L., Hughes M., Kondo, M., Rebel, V.I., Schmucker D. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* **309:**1874–1878.

Wei C.L., Ng P., Chiu K.P., Wong C.H., Ang C.C., Lipovich L., Liu E.T., Ruan Y. 2004. 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl. Acad. Sci. USA* **101:**11701–11706.

Wheelan S.J., Church D.M., Ostell J.M. 2001. Spidey: A tool for mRNA-to-genomic alignments. *Genome Res*. **11:**1952–1957.

Wiles M., Andreassen P. 2006. Monoclonals: The billion dollar molecules of the future. *Drug Discov. World* **4:**17–23.

Wilton S.D., Fletcher S. 2005. RNA splicing manipulation: strategies to modify gene expression for a variety of therapeutic outcomes. *Curr. Gene. Ther*. **5:**467–483.

Xia H., Bi J., Li Y. 2006. Identification of alternative 5'/3' splice sites based on the mechanism of splice site competition. *Nucleic. Acids. Res.* **34:**6305–6313.

Xie H., Zhu W.Y., Wasserman A., Grebinskiy V., Olson A., Mintz L. 2002. Computational analysis of alternative splicing using EST tissue information. *Genomics* **80:**326–330.

Xing Y., Lee C.J. 2005. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet*. **1:**e34.

Xing Y., Xu Q., Lee C. 2003. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett.* **555:**572–578.

Xu Q., Lee C. 2003. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic. Acids. Res.* **31:**5635–5643.

Yeo G., Holste D., Kreiman G., Burge C.B. 2004. Variation in alternative splicing across human tissues. *Genome. Biol*. **5:**R74.

Yeo G.W. 2005. Splicing regulators: Targets and drugs. *Genome Biol*. **6:**240.

Yeo G.W., Van Nostrand E., Holste D., Poggio T., Burge C.B. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci. USA* **102:**2850–2855.

Yura K., Shionyu M., Hagino K., Hijikata A., Hirashima Y., Nakahara T., Eguchi T., Shinoda K., Yamaguchi A., Takahashi K., Itoh T., Imanishi T., Gojobori T., Go M. 2006. Alternative splicing in human transcriptome: Functional and structural influence on proteins. *Gene.* **380:**63–71.

Zavolan M., Kondo S., Schonbach C., Adachi J., Hume D.A., Hayashizaki Y., Gaasterland T. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res*. **13:**1290–1300.

Zhang C., Li H.R., Fan J.B., Wang-Rodriguez J, Downs T., Fu X.D., Zhang M.Q. 2006. Profiling alternatively spliced mRNA isoforms for prostate cancer classification. *BMC Bioinformatics* **7:**202.

Zhang H., Hu J., Recce M., Tian B. 2005. PolyA_DB: A database for mammalian mRNA polyadenylation. *Nucleic. Acids. Res.* **33:**D116–120.

Zhang X.H., Chasin L.A. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*. **18:**1241–1250.

Zhang X.H., Leslie C.S., Chasin L.A. 2005. Computational searches for splicing signals. *Methods* **37:**292–305.

Zheng C.L., Fu X.D., Gribskov M. 2005. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11:**1777–1787.

Zheng C.L., Kwon Y.S., Li H.R., Zhang K., Coutinho-Mansfield G. 2005. MAASE: An alternative splicing database designed for supporting splicing microarray applications. *RNA* **11:**1767–1776.

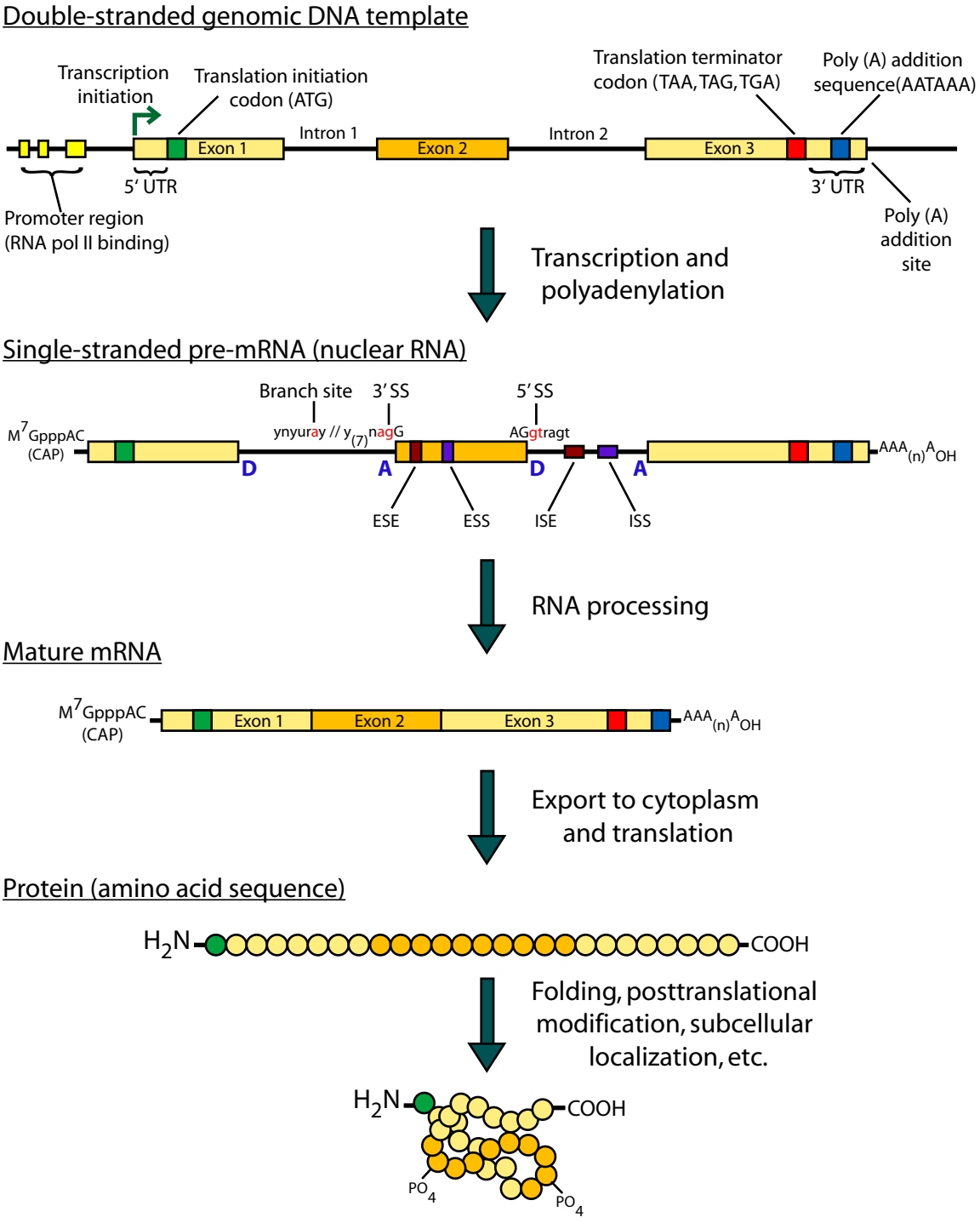

Plate I. Gene transcription and RNA processing. Expression of a typical protein-coding gene involves: gene transcription, pre-mRNA processing and polyadenylation. Each of these processes is regulated by components of the transcription machinery, which recognize sequence motifs in the DNA template and pre-mRNA molecule. After pre-mRNA processing, the mRNA is exported to the cytoplasm where ribosomes translate it into protein. Abbreviations: (UTR) untranslated region; (D) donor site; (A) acceptor site; (SS) splice site; (ESE) exonic splicing enhancer; (ESS) exonic splicing silencer; (ISE) intronic splicing enhancer; (ISS) intronic splicing silencer.

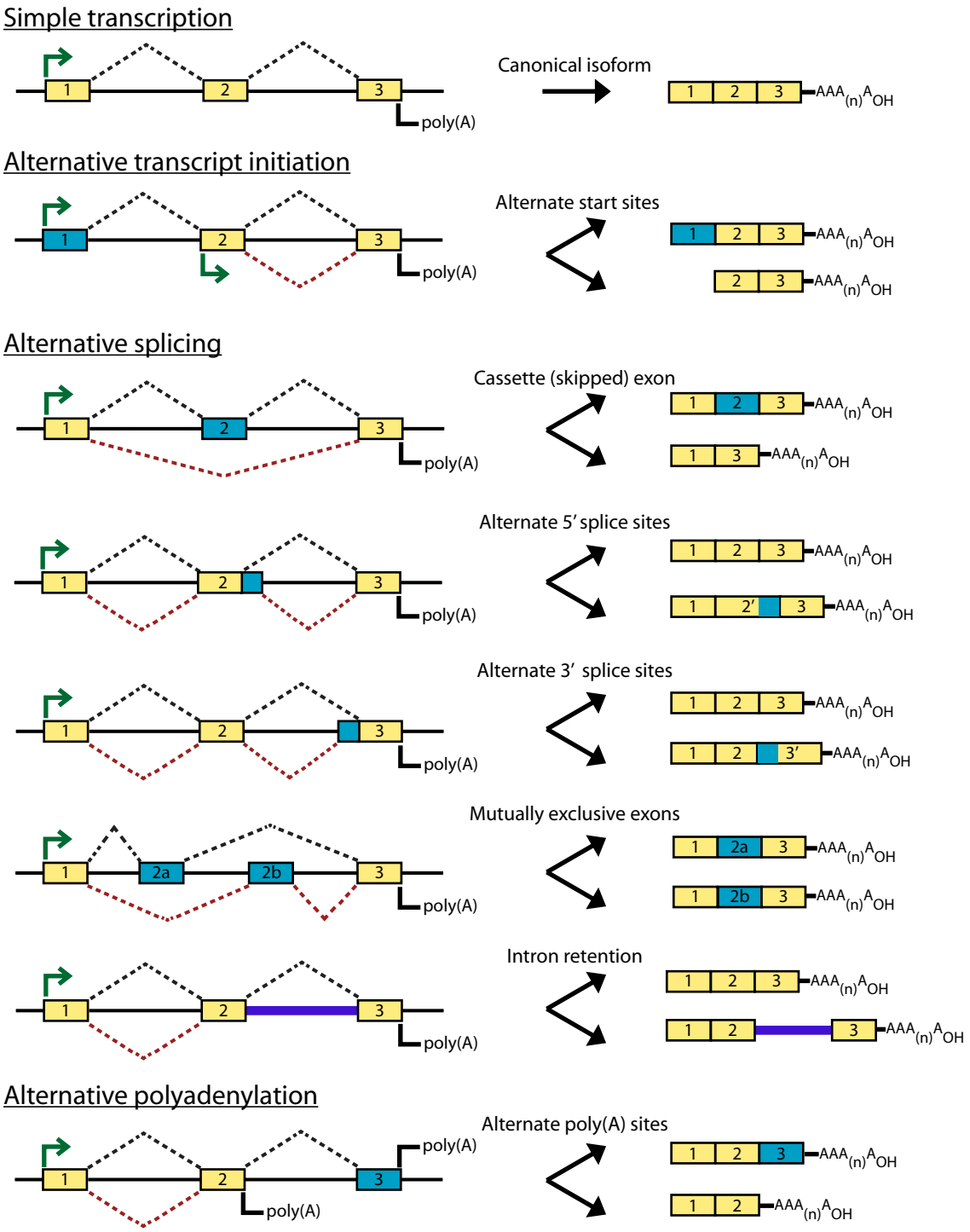

Plate II. Types of alternative transcription (AT). Gene models are depicted as exons (colored rectangles) connected by introns (black lines). Green arrows indicate transcription initiation sites, dotted lines indicate splicing patterns and polyadenylation sites are denoted as 'poly (A)'. The mRNA products generated by each type of AT are shown to the right of each gene model. Simple transcription is contrasted with alternative transcript initiation, the five major classes of alternative splicing, and alternative polyadenylation. In each model, yellow exons are constitutive and blue exons are alternative.

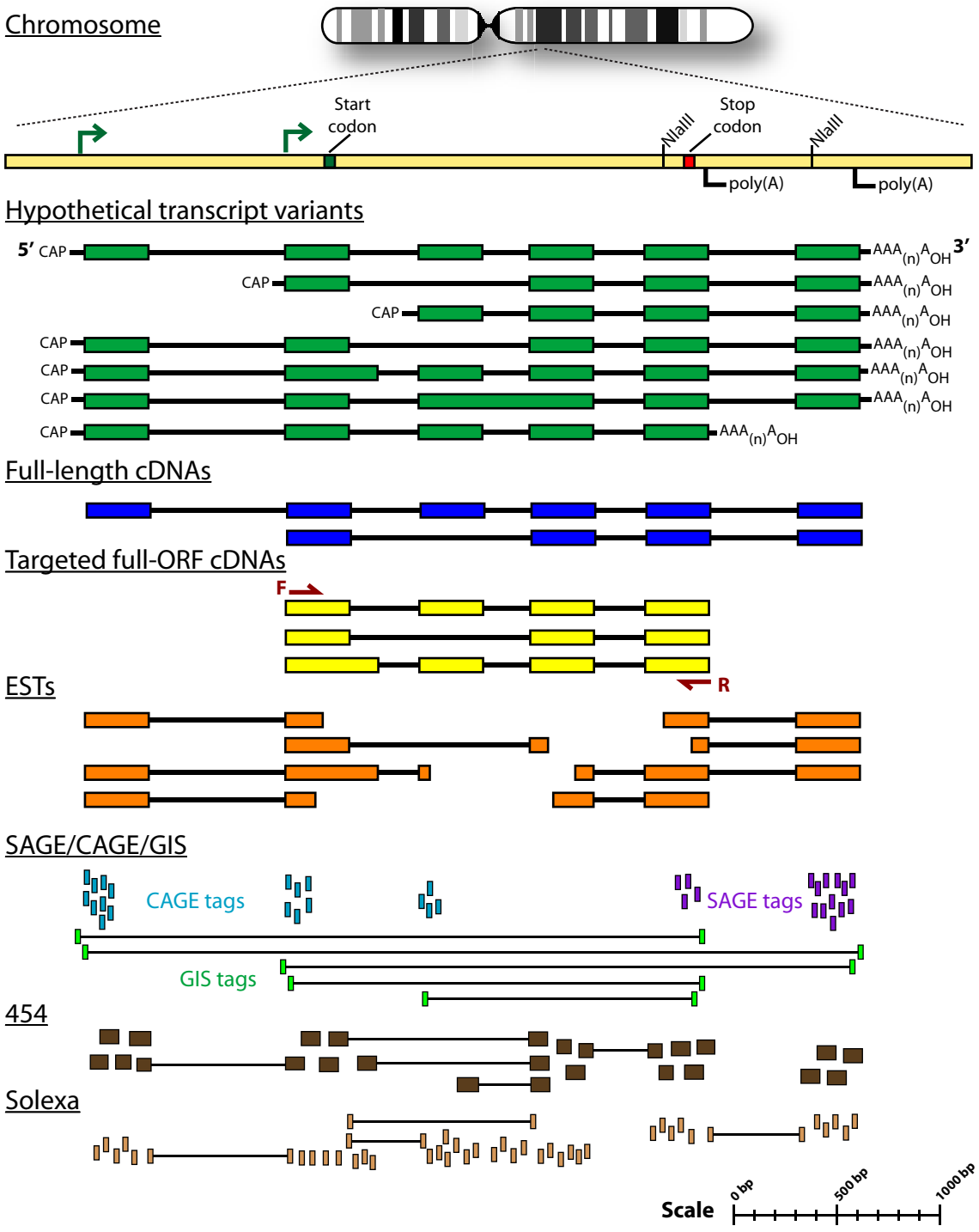

Plate III. Sequence-based methods for profiling transcript diversity. Hypothetical transcript sequences consisting of exons (green rectangles) with intervening introns (black lines) are depicted as gapped alignments to a reference genome. The following tracks represent sequences generated by each sequence based method. Human genes have an average of 10 exons with an average length of 250 bp. The methods are displayed in order of least to most quantitative. Abbreviations: (EST) expressed sequence tag; (SAGE) serial analysis of gene expression; (CAGE) capped analysis of gene expression; (GIS) gene identification signature.

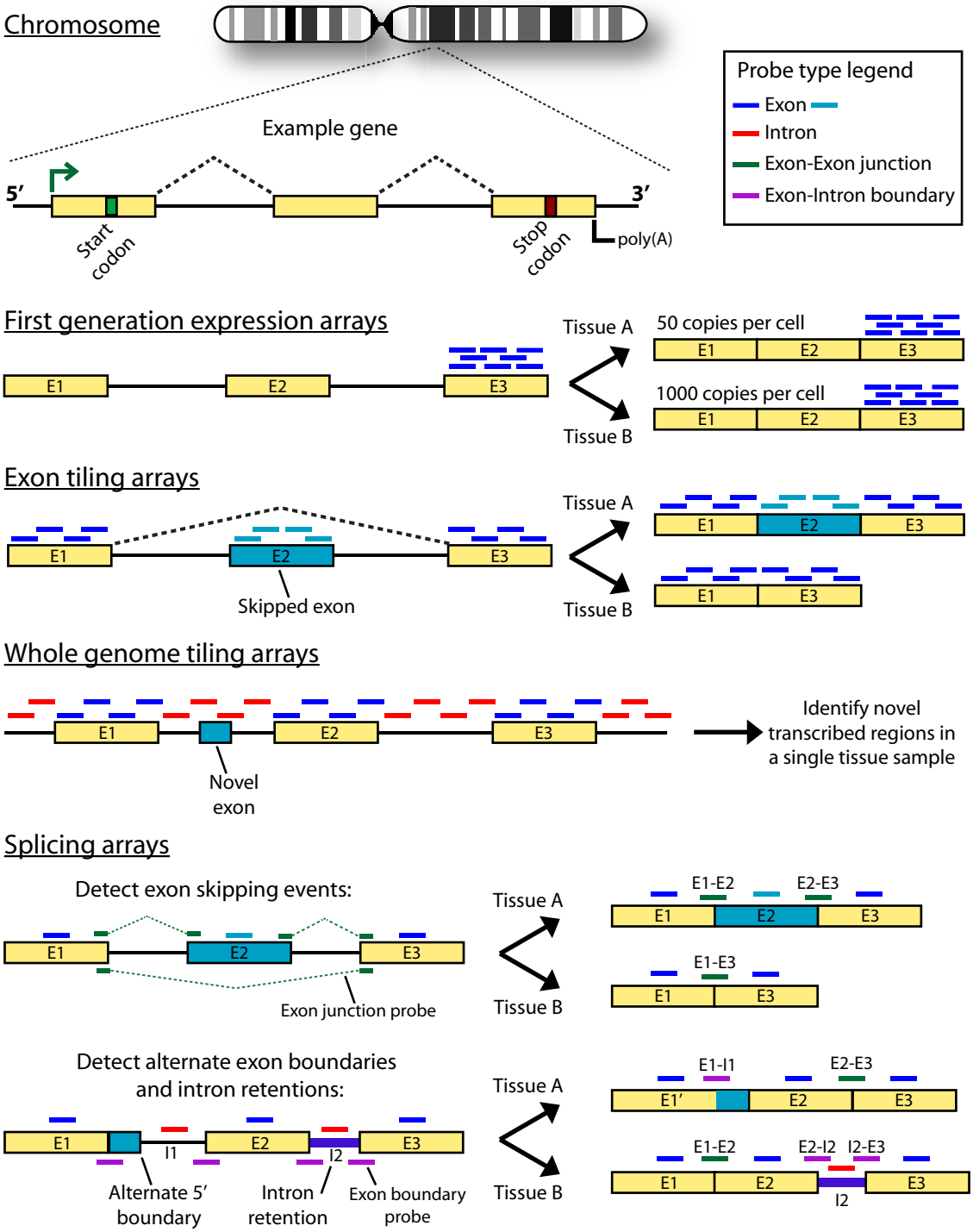

Plate IV. Microarray based methods for profiling transcript diversity. Gene models are depicted as exons (colored rectangles) connected by introns (black lines). Hypothetical differences in mRNA products, which can be detected by each array method, are depicted to the right of each gene model. In each model, yellow exons are constitutive and blue exons are alternative. Differences in array design strategy, particularly the position and types of oligonucleotide probes used are shown above each gene model as colored horizontal lines.