

## Detection and measurement of alternative splicing using splicing-sensitive microarrays

Karpagam Srinivasan <sup>a,b</sup>, Lily Shiue <sup>a</sup>, Justin D. Hayes <sup>a,b</sup>, Ross Centers <sup>b</sup>, Sean Fitzwater <sup>b</sup>, Rebecca Loewen <sup>b</sup>, Lillian R. Edmondson <sup>b</sup>, Jessica Bryant <sup>b</sup>, Michael Smith <sup>b</sup>, Claire Rommelfanger <sup>b</sup>, Valerie Welch <sup>a</sup>, Tyson A. Clark <sup>a</sup>, Charles W. Sugnet <sup>c</sup>, Kenneth J. Howe <sup>a</sup>, Yael Mandel-Gutfreund <sup>a,b</sup>, Manuel Ares Jr. <sup>a,b,\*</sup>

<sup>a</sup> Department of Molecular, Cell, and Developmental Biology, Center for Molecular Biology of RNA, Sinsheimer Laboratories, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

<sup>b</sup> Hughes Undergraduate Research Laboratory, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

<sup>c</sup> Department of Biomolecular Sciences and Engineering, Baskin School of Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

Accepted 22 September 2005

### Abstract

Splicing and alternative splicing are major processes in the interpretation and expression of genetic information for metazoan organisms. The study of splicing is moving from focused attention on the regulatory mechanisms of a selected set of paradigmatic alternative splicing events to questions of global integration of splicing regulation with genome and cell function. For this reason, parallel methods for detecting and measuring alternative splicing are necessary. We have adapted the splicing-sensitive oligonucleotide microarrays used to estimate splicing efficiency in yeast to the study of alternative splicing in vertebrate cells and tissues. We use gene models incorporating knowledge about splicing to design oligonucleotides specific for discriminating alternatively spliced mRNAs from each other. Here we present the main strategies for design, application, and analysis of spotted oligonucleotide arrays for detection and measurement of alternative splicing. We demonstrate these strategies using a two-intron yeast gene that has been altered to produce different amounts of alternatively spliced RNAs, as well as by profiling alternative splicing in NCI 60 cancer cell lines.

© 2005 Elsevier Inc. All rights reserved.

**Keywords:** Alternative splicing; Microarrays; Oligonucleotides; Gene annotation; Cancer

### 1. Introduction

Splicing is required for proper expression of the vast majority of eukaryotic genes. Without information about splicing, prediction of protein coding potential of a eukaryotic genome is difficult if not impossible [1]. Compounding the problem, different mRNAs from the same gene (mRNA isoforms or splice variants) can be produced, encoding related proteins with distinct functions [2,3]. Alternative splicing is a major source of protein diversity in higher

eukaryotes, expanding the coding potential of the genome [2,4–6]. There are few methods available to most researchers that allow large-scale changes in splicing of many genes to be detected or measured in a single experiment.

The use of microarrays has transformed the way gene expression is studied in a wide variety of eukaryotic systems. Microarrays can screen for the presence of mRNA from many genes in many defined samples, allowing the experimentalist to match changes in the landscape of gene activity to other biological events [7–9]. Unfortunately, standard microarray designs that strive to measure only the overall level of mRNA from each gene provide little or no information about splicing. Standard methods of splicing analysis such as reverse transcriptase primer extension,

\* Corresponding author. Fax: +1 831 459 3737.

E-mail address: [ares@biology.ucsc.edu](mailto:ares@biology.ucsc.edu) (M. Ares Jr.).

nuclease protection mapping, and cDNA cloning are too laborious to scale up for analysis of large numbers of genes and samples. To more generally determine the contribution of alternative splicing to programs of gene expression, parallel methods for measuring splicing en masse have been developed.

Several groups have contributed to the development of parallel methods for analyzing splicing. One method is to use arrays of spotted oligonucleotides designed to discriminate different RNAs from the same gene by targeting the distinct sequence features that characterize their differential splicing. Oligonucleotide probes that span splice junctions sample the presence and amount of RNA joined by particular splicing events. Other oligonucleotide probes are placed within exons in order to determine individual exon levels as well as overall mRNA level [10,11]. Recent efforts using both the Affymetrix platform [12,13] and the Rosetta platform [14–18] have shown promise in the application of this approach to alternative splicing. A second method for detecting the joining of specific exons provides for parallel measurement of many splicing events in the same experiment by using the spliced RNA as a template for the ligation of oligonucleotides. Ligation of specific pairs of oligonucleotides creates a PCR replicon that is detected after amplification, and reveals which exons are joined with which in the sample [19].

Here we describe the design, application, and analysis of simple printed oligonucleotide microarrays to the parallel detection and measurement of alternative splicing. The first section of this article discusses theoretical aspects of the problem. In the second section, we present two simple contexts in which to test and illuminate the principles and issues. First, we use a modified two-intron yeast gene altered in the branch point of the first intron that gives rise to different amounts of alternatively spliced RNAs to show that specificity of the array elements for a given isoform depends as expected on oligonucleotide probe length and hybridization temperature. Second, we identify and validate differential alternative splicing in human cancer cell lines using a small array carrying probes for 64 genes. Well-designed splicing-sensitive microarrays can be broadly applied in small scale or on genomic scale for parallel detection of alternative splicing in a flexible, cost-effective fashion.

## 2. Theoretical considerations for splicing sensitive microarrays

### 2.1. Annotation of alternative splicing for microarray design and analysis

#### 2.1.1. Alternative splicing and gene models

To detect and measure alternatively processed RNAs requires a hypothesis or model describing the structures of the RNAs that need to be distinguished. Alignment of the sequences of mature RNAs to the gene sequence allows identification of exon boundaries and splice junctions. A

common result is called exon skipping, wherein two alternative mRNA forms are produced that differ by the presence of a single exon (Fig. 1A). Since alternative mRNAs are formed, alternative exons and junctions (or their predictions) must be defined. We use the term *gene model* to describe the set of alternative RNAs that might be produced from a gene. The gene model specifies two main features of gene expression: (1) all genomic regions ever included in any RNA from the gene, represented as *spans*, and (2) all *paths* through the gene model that lead to production of any of the RNAs of interest. A path consists of a “start,” where transcription is considered to begin, a series of splice site joining events whereby specific 5' splice sites are joined to specific downstream 3' splice sites in a 5' to 3' sequence, followed by an “end,” where the transcript is considered to terminate. Fig. 1B shows a generalized gene model describing a single cassette exon, which would be developed from the biological observations from Fig. 1A. Due to alternative splicing, gene models can become quite complex quite rapidly. A next level of complexity is shown in Fig. 1C, in which two alternative exons are found between constitutive exons. One model shows a “double cassette” mode of splicing in which either or both cassettes can be included or skipped, whereas the other model shows “mutually exclusive exons” in which only one or the other exon is included.

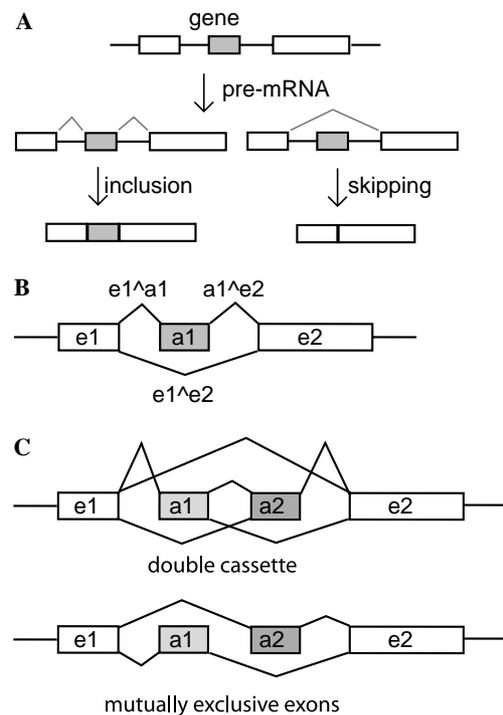


Fig. 1. Splicing observations and gene models. (A) A simplified eukaryotic gene is encoded in the genome discontinuously and primary transcripts derived from it can be spliced to include or skip an internal exon. (B) A gene model for the gene in (A) labeled to show constitutive exons e1 and e2, the alternative exon a1, and the three splice junctions e1<sup>a1</sup>, a1<sup>e2</sup>, and e1<sup>e2</sup>. (C) Two gene models that differ by splice junction utilization, the double cassette (above) and mutually exclusive exons (below).

Gene models can be derived from automated alignment of cDNAs to genomic sequence by any of a number of methods [20–24], provided these are judiciously filtered to ensure that true splice sites border each internal exon. Other knowledge about splice site joining, such as sequences of PCR products or nuclease protection experiments, must often be incorporated to make the best gene models for genes of interest that are poorly represented by EST or mRNA database entries. Ever since alternative splicing was discovered, eukaryotic genes have been represented as shown in Fig. 1B. To keep track of large amounts of splicing information, it is useful to formalize the application of these traditions. Careful annotation of the target genes is critical for designing and analyzing a useful, cost-effective array.

### 2.1.2. Recording gene models in the computer

To translate splicing biology into terms that can be used in classification and computation, it is important that specific terms have precise meanings. Although casual usage among our colleagues may differ, for communication with computers we have started using the following terms. We define an *alternative splicing event* to exist if a single 5' (or 3') splice site is joined to at least two different 3' (or 5') splice sites. In the simplest cases, an event can be described by naming the three splice sites involved (see below). The term *mode* of alternative splicing describes the type or combination of types of alternative splicing events observed, for example, an “alternative 5' splice site” would be one of the

simplest modes, and “mutually exclusive exons” would be a more complex mode. For array analysis of very complex modes it is useful to define a *region* of alternative splicing, within which only one mode occurs (for example a simple cassette), but which may lie within a part of the gene that can be skipped entirely. Identifying constitutive regions of a transcript, that is those parts that are present in every transcript from the gene, can be difficult when a gene has many complex modes of splicing. In these cases the somewhat oxymoronic concept of *locally constitutive regions* can be used to design array elements that allow measurement of simpler local modes of splicing without concern for skipping of a larger region.

For annotating and archiving splicing events, we reference splicing events to a genome sequence, using genomic coordinates from specified assembled versions of the appropriate genome found at the UCSC Genome Browser site (<http://genome.ucsc.edu>). Splicing data is captured by EST and mRNA alignment to the genome and a “splicing graph” is obtained (see Fig. 2, [25–29]). In our system, the boundaries of the gene model on the genome are called an *annotation*, which consists of the genomic coordinates that entirely contain the gene model and the strand on which it resides. A *span* is a pair of genomic coordinates that describe genome positions that form a part of the gene model relevant to splicing. Spans are classified by *span type*, which tells whether the span describes an exon (0), splice junction (2), intron (1), start (S), etc.

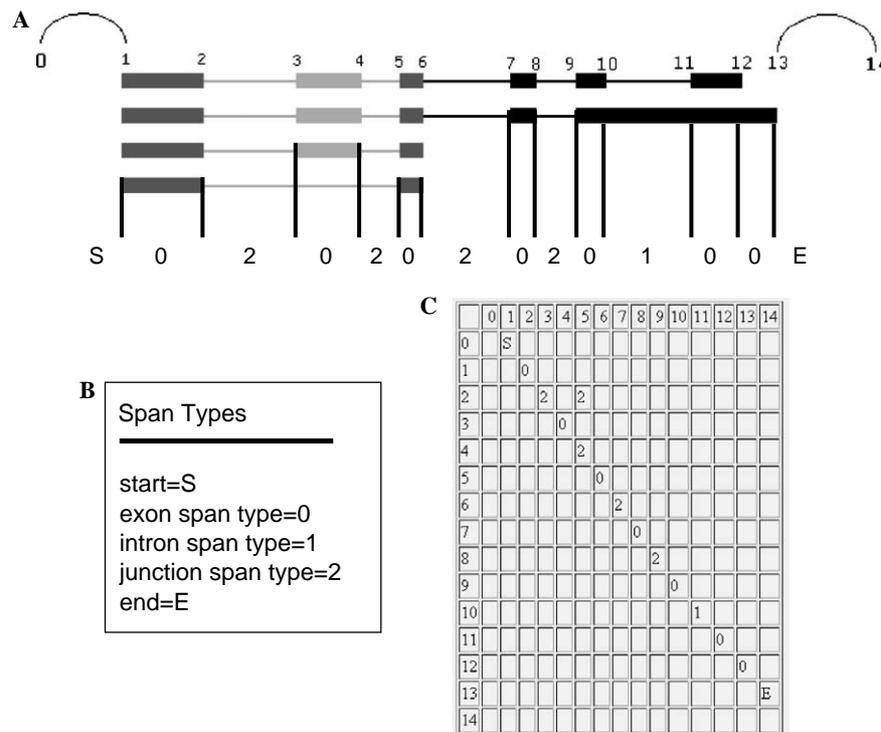


Fig. 2. A method of encoding splicing information in a graph. (A) An annotation in the genome of a gene with alternative splicing. Spans (bounded by vertical bars) describe where the gene starts and ends and where the exons and splice junctions are in genomic coordinates (labeled 1–14). Span type is labeled below. (B) Examples of different span types. (C) A matrix representation of the gene shown in (A) indicating the location and types of spans in the gene model. Paths can be traced through the gene in order to generate virtual isoforms that represent possible mRNAs described by the gene model.

A gene model displayed in terms the computer might understand is shown in Fig. 2. An annotation begins with a start span (S), and each position in the model is given a number that represents the different vertices of the graph, with the spans representing the edges (Fig. 2C, for a discussion of graph theory and splicing, see [26–28]). The next span after the start span in this model is an exon span (type 0), after which comes a splice junction (span type 2) that joins an alternative exon (positions 2–3). This is indicated by a 2 in row 2, column 3 of the graph. Skipping of this exon involves joining position 2 to position 5 by a splice junction, indicated by the 2 in row 2, column 5 in the graph (Fig. 2C). Thus at position 2 the path diverges, and alternative splicing is indicated. The inclusion path goes through an exon span (the 0 in row 3, column 4) and then another splice junction span (the 2 in row 4, column 5) at which point the paths converge again. Although this language may seem cumbersome and obvious, the strict specification of these events is critical for summarizing and analyzing large amounts of splicing data. In simple terms, the alternative splicing regions of the transcript can be identified as those places in the gene model between points where the path diverges and converges, whereas the constitutive regions of the transcript are those places in the gene model where there is only one path.

### 2.1.3. The challenge of capturing complex modes of splicing

There are many modes of splicing, simple and complex [2,3]. The complex modes appear to be built up by the combination of simple modes of splicing. For example, the combination of the simple modes of cassette exon and an alternative 5' splice site can be classified as a more complex mode. Adding the possibility of alternative promoter (starts) and polyadenylation sites (ends) greatly increases the number of possible modes that can be envisioned. It is not yet clear how valuable this classification might be in terms of understanding the mechanisms of splicing regulation. For the application of arrays to splicing it can be valuable to restrict analysis to simple modes such as cassette exons or alternative 5' or 3' splice sites. Overall, it is more important to identify first which junctions and parts of exons in the gene model are alternative, and which are constitutive. Then as the analysis proceeds, this information can be assembled to investigate more complex relationships.

## 2.2. Designing arrays

### 2.2.1. What is the intended purpose of the array?

The goals of the investigation should influence the design of the array. Our first application was to measure the simplest alternative (to splice or not to splice) for introns in the yeast genome [11]. This is an example of the “annotate to design” approach, in which splicing is exhaustively annotated [30,31] to design an array that will provide expression information for the known RNA forms of interest, to profile gene expression at splicing-level resolution. In applying

this to complex vertebrate patterns of alternative splicing it becomes even more important to capture as much information as possible to develop the best gene models. There are many, many databases available containing curated alternative splicing information based on EST/mRNA clustering and alignments. Many of these events are referenced to clusters of ESTs and it may take some work to map the splicing events they contain to genomic coordinates. Arrays designed using comprehensive annotation are best for estimating expression of mRNAs of known structure, but may be less useful for the systematic discovery of new isoforms.

If the goal is to discover new alternative splicing, gene models with predicted but unsubstantiated elements can be created, and array features can be designed to ask whether or not RNA with a particular structure is expressed. Then array data can be used to substantiate or refute aspects of the gene model. For example one group [32] simply made probes for many genomic regions (genomic tiling) and used their signals to distinguish introns from exons. Another used RefSeq annotations, which at the time were largely devoid of alternative splicing information, to make arrays of splice junctions to search for cases in which signals for consecutive junctions were lost, indicating exon skipping [17]. These are examples of the “design to annotate” approach, in which the goal is to discover new exons and new examples of alternative exons to annotate gene structure and function with splicing level resolution.

### 2.2.2. Design of array elements for each gene model

Once the gene model is settled, array probes are designed to capture target RNA regions in the sample that are common to all transcripts (constitutive regions), as well as those specific to one isoform or another (alternative regions). We consider many possible target sequences to capture, including exons (E), introns (I), exon–intron (EI) junctions (i.e., 5' splice sites), intron–exon (IE) junctions (i.e., 3' splice sites) and exon–exon junctions (splice junctions, SJ). The first consideration is to obtain measurements from the constitutive regions of RNA to estimate the expression level of the gene in question. Although it may seem obvious that alternative splicing cannot be detected from genes that are not expressed, it is very important to note the error potential in the measurement of alternative splicing for genes whose expression levels are near the limit of detection. In addition, the measurement of the total RNA level is useful in comparing changes in the amount of each of the different alternative regions by creating “splicing indexes” derived by within-gene normalization [11]. For this purpose, probes to the constitutive regions of the transcript (e.g., e1 and e2 in Fig. 1B) are critical.

After planning to estimate total gene-derived RNA levels using constitutive probes, analysis of probes specific for alternative splicing events are needed. Capturing information about alternative splicing requires designing probes for two additional classes of measurements: *exon representation*, which estimates the levels of individual exons or specific variant forms of exons, and *splice junction representation*, which

estimates how much of each splice junction sequence is present. Theoretically, splice junctions have the most information, however in practice obtaining reliable signals can be challenging (see below). Thus, both exon and splice junction representation for alternative regions of the transcripts are necessary to make the best estimates of the structure and composition of the transcript pool.

To visualize how this works, consider the common simple case (Fig. 1B) where a single cassette exon ( $a_1$ ) is positioned between two constitutive exons ( $e_1$  and  $e_2$ ). There are three exons ( $e_1$ ,  $a_1$ , and  $e_2$ ), and three splice junctions of interest, the two on either side of  $a_1$  that join to the constitutive exons on either side (“include” junctions  $e_1\hat{a}_1$  and  $a_1\hat{e}_2$ ), as well as the junction formed between  $e_1$  and  $e_2$  when  $a_1$  is skipped (the skip junction,  $e_1\hat{e}_2$ ). From our gene model, we expect the level of RNA for  $e_1$  to be the same as  $e_2$ , since every transcript has both exons. Thus one expectation is that  $e_1 \approx e_2$ . In addition, since the following probes all signal inclusion of  $a_1$ , we expect  $e_1\hat{a}_1 \approx a_1\hat{e}_2 \approx a_1$ . When inclusion is complete (no skipping of  $a_1$ ),  $e_1 \approx a_1 \approx e_2 \approx e_1\hat{a}_1 \approx a_1\hat{e}_2$ , and  $e_1\hat{e}_2$  is undetectable. As skipping increases,  $a_1$  decreases such that  $e_1 \approx e_2 > a_1$ , and  $e_1\hat{e}_2$  increases. When skipping is complete,  $e_1 \approx e_2 \approx e_1\hat{e}_2$ , and  $a_1$ ,  $e_1\hat{a}_1$ , and  $a_1\hat{e}_2$  are undetectable. These predictions are based on the gene model and assume (1) that all oligos perform equally in capture of the target, and (2) that each region of the target RNA is equivalently represented in the labeling reaction. Since these last two assumptions are rarely valid in practice, the expectations above serve only as an initial guide.

It might seem that with good estimates of exon representation, splice junction representation would be redundant or unnecessary. Unfortunately, distinguishing more complex modes, such as those in Fig. 1C, cannot be done using only exon representation data. Consider a double cassette exon model, which results in the production of four possible forms (neither  $a_1$  nor  $a_2$ ,  $a_1$  only,  $a_2$  only, both  $a_1$  and  $a_2$ , Fig. 1C, top), and a strictly mutually exclusive exons model, which results in production of only two forms ( $a_1$  only,  $a_2$  only, Fig. 1C, bottom). Knowing only that exon  $a_1$  and  $a_2$  are each present in 50% of the transcripts does not help distinguish between these two models. Estimates of how often both or neither exon is included in the same transcript are necessary, and these rely on sampling the two splice junctions unique to the double cassette exon model. Thus array elements designed to sample the  $e_1\hat{e}_2$  (skipping both  $a_1$  and  $a_2$ ) and  $a_1\hat{a}_2$  (include both  $a_1$  and  $a_2$ ) junctions (Fig. 1C) are needed. The more complex the splicing mode, the more probes must be designed to describe it.

### 2.2.3. Biochemical considerations for probe design

The ability to distinguish differently spliced transcripts arising from the same gene depends on the specificity of the oligonucleotide probes for their targets. Since all probes are hybridized under the same conditions as part of an array, similarity in predicted melting temperature ( $T_m$ ) is desir-

able. It is wise to avoid oligonucleotides that have internal secondary structure or low information content. Finally, comparison of the chosen sequence back to the genome to detect cross-hybridization is valuable, so that sequences similar to highly transcribed repeats or shared sequences of paralogous genes can be avoided where possible. Numerous methods and heuristics for choosing oligonucleotide sequences for microarrays are available (e.g., 33–35). Obviously where the RNA region to be detected is large, for example in a large exon, many choices for a probe sequence are available. Where the target sequence is highly constrained, as for a small exon or a splice junction, the choice of probe sequences is limited. Thus the chances of obtaining a probe with optimum characteristics for hybridization are less at the junctions than they are within large exons.

Predicting how well a chosen DNA sequence will behave as an array probe is not straightforward [36]. Since the oligonucleotide probe is in excess near the surface of the array, annealing between the target and the probe is a pseudo-first order reaction whose rate should be approximately dependent on the target concentration. Hybridization efficiency is only partly dependent on the predicted melting temperature ( $T_m$ ) of the duplex and the temperature at which the hybridization is done, traditionally at  $T_m - 20^\circ\text{C}$  or  $T_m - 25^\circ\text{C}$  [37]. Using  $T_m$  to predict array probe efficiency is limited by the fact that the array operates by annealing, rather than melting. During annealing, the structure of the individual strands must be overcome to form a duplex, whereas the melting reaction has no such associated costs. For some oligonucleotides this is a negligible consideration, for others it is practically insurmountable, greatly influencing the performance of array elements. In addition, depending on the attachment chemistry there may be surface effects that prevent the part of the oligonucleotide sequence nearest the array surface from participating in collisions necessary for annealing. Finally, the overall length of the probe may be expected to influence how much signal is captured, with longer oligos more likely to give stronger signals in general. These theoretical uncertainties mean that despite the best efforts to choose the best probe, certain target regions may remain recalcitrant to detection and accurate measurement by this method.

### 2.2.4. The problem of half-junction crosstalk

Splice junction probes present a unique problem. Since about half of a splice junction probe will be derived from one exon and about half from another, each junction probe has perfect complementarity over about half of its length to other RNA forms that contain a different exon [11,13,16]. For example in Fig. 1B, the  $e_1\hat{e}_2$  junction found in the exon-skipped isoform shares half-junctions with the other junctions  $e_1\hat{a}_1$  and  $a_1\hat{e}_2$ , found in the exon-included isoform. If the hybridization stringency is too low, the  $e_1\hat{e}_2$  (skip) probe will capture exon-included RNA sequences through hybridization to either of its half-junction sequences. Under these conditions, the true signal from the exon-skipped RNA will be partly obscured by cross

hybridization to the exon-included form. Since alternative RNA forms commonly share half-junctions, half-junction crosstalk must be minimized. The most specific junction probes will be those that detect poorly any sequence that matches perfectly over half of its length, but detect well the sequence that matches perfectly over its whole length.

To estimate biochemical parameters for half-junction cross-hybridization for probes of different length, one can compare the difference in theoretical melting temperature between the average full-length oligonucleotide and either of its two halves ( $\Delta T_m$ , Fig. 3), using standard methods of estimating  $T_m$  [38]. A large difference between melting temperatures ( $T_m$  full –  $T_m$  half-junction =  $\Delta T_m$ ) should afford increased specificity for the true junction. As sequences get longer,  $T_m$  increases steeply and then plateaus [38,39], such that the difference in predicted melting temperature between a sequence and one of its half-junction sequences ( $\Delta T_m$ ) decreases (Fig. 3A). This means that specificity will likely decrease with increasing oligo length. In a calculation of average predicted  $T_m$  for random sequences with yeast-like base compositions, the  $\Delta T_m$  for an average full-length 30mer and its half 15mer is 14.0°C, while the  $\Delta T_m$  for the

average 50mer and 25mer is only 8.2°C (Fig. 3A). Therefore, to a first approximation, shorter oligos should have reduced half-junction cross-hybridization and should provide an advantage for distinguishing different mRNA isoforms. If the oligo is too short however, it may not capture sufficient target to give a robust signal. This tradeoff between sensitivity and specificity is at the crux of obtaining and analyzing splicing microarray data.

Even more severe problems with half-junction crosstalk can occur where the sequences of the alternatively joined exons match past the splice junction, in effect extending the sequence similarity between the probe's intended target and the alternatively spliced target. For example, the 5' end of the alternative exon 25 of human ITGA6 matches the 5' end of exon 26 at 12 of their first 14 bases, such that the e1^a1 junction is very similar to the e1^e2 junction (Fig. 3B). This situation occurs frequently but not usually to this extent. Since the 5' splice site consensus for vertebrates include the sequence (C/A)AG in the exon upstream, there are many junctions that match at least a few bases past the site of exon joining. We are currently exploring probe design methods to address such cases. Below we test some of the expectations from the theoretical discussion above.

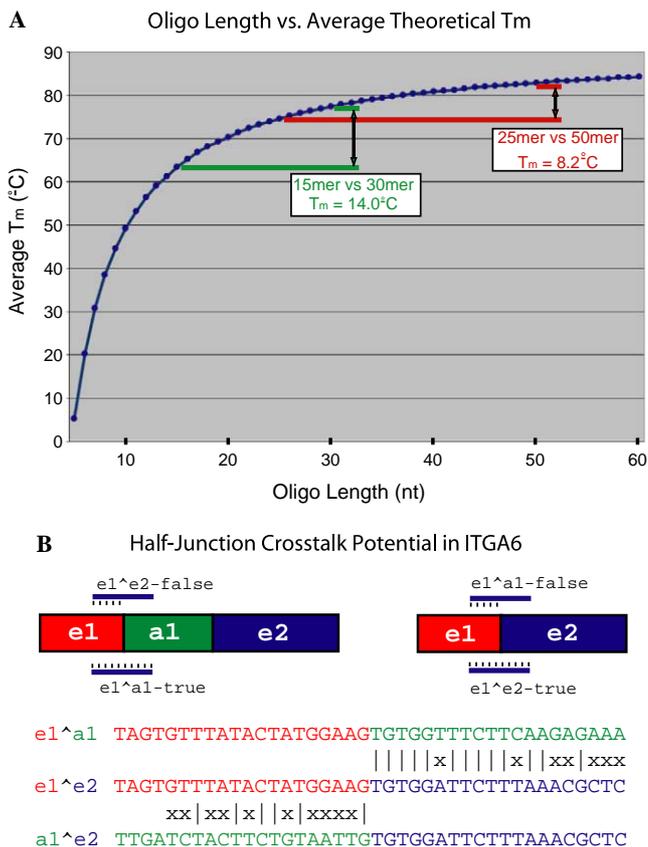


Fig. 3. The half-junction cross talk problem. (A) Relationship of oligo length to calculated  $T_m$ . As the oligo gets longer, its potential in discriminating other isoforms that share its half-junction decreases. (B) An example of severe potential for half-junction cross talk due to extended similarity between two alternative exons (a1 and e2) just downstream of their 3' splice sites. No such similarity exists between the regions of e1 and a1 just upstream of their 5' splice sites. |= base match, × = base mismatch.

### 3. Application and practical considerations for splicing-sensitive microarrays

#### 3.1. Testing oligonucleotide performance using the yeast *DYN2* gene

##### 3.1.1. A modified *DYN2* test gene

The *DYN2* gene is one of several known multi-intron genes in the genome of *Saccharomyces cerevisiae*. Normally, the second exon of *DYN2* is included, and only a very small amount of mRNA that skips this exon can be detected [40]. Mutations in or near the first intron branchpoint stimulate skipping of the second exon and to a lesser extent allow retention of the first intron (Fig. 4A [40]). Severe branchpoint mutations prevent formation of the exon 2 included form, converting the mRNA population into a mixture of exon 2 skipped and intron 1-retained transcripts (Fig. 4A). These observations led us to use *DYN2* as a test bed for demonstrating the concept of measuring alternative splicing with microarrays, because we could create yeast strains that differentially alternatively splice *DYN2* transcripts.

##### 3.1.2. Array design to test probes for measuring *DYN2* alternative splicing

We developed a gene model for *DYN2* (Fig. 4B) and designed array elements to specifically detect the three spliced products, as well as the unspliced transcript and the incompletely spliced intermediates. For most practical applications to alternative splicing, detection of precursors and intermediates may not be necessary. Since we wanted to obtain an exhaustive description of splicing for this one gene, we wanted probes for all possible exons, introns, and

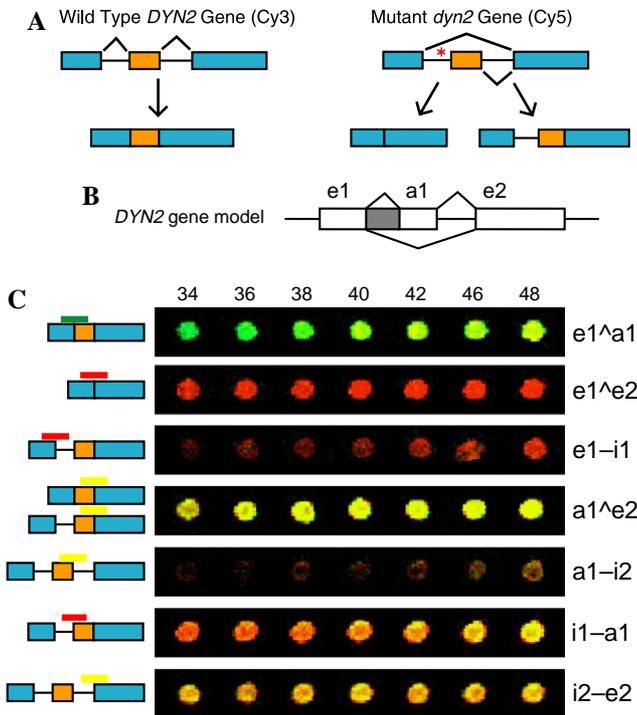


Fig. 4. Test of specificity for oligo probes of different lengths using *DYN2* from yeast. (A) Observations on splicing in *dyn2* intron 1 branchpoint mutants. (B) A gene model for *DYN2* and its mutants incorporating the observations in (A). (C) False color images of spots scanned after hybridization at 65 °C. Sizes of the centered oligos are shown at top, the junctions represented are indicated at right, and a cartoon of the target and its probes expected behavior is shown at left. The wild type *DYN2* sample was labeled with Cy3 (green) and the *dyn2* mutant sample was labeled with Cy5 (red). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

junctions (Fig. 4C). The main goal of our experiment was to study the effect of probe length on specificity and sensitivity of signal capture where the amounts of alternative RNA forms could be well-quantitated by standard means. Given this, we settled on a set of oligonucleotide probes of varying length to use as array elements. Because we reverse transcribe the RNA into a Cy3 or Cy5 labeled cDNA to create the labeled target sequences, we want our array probes to represent the strand able to capture the labeled cDNA, i.e., the same strand as the RNA. Other labeling protocols may demand the use of the other strand on the array, and this must be carefully considered during the design stage.

To sample effects at the junctions, we ordered a series of oligos centered across the junctions that were 34, 36, 38, 40, 42, 44, 46, and 48 nucleotides long. We also ordered sets of oligos of different lengths for exons and introns of *DYN2*, as well as control probes for other yeast mRNAs and other RNAs that could be “spiked in” to the sample to control for labeling efficiency. All oligos were ordered with 5' amino linkers in order that they could covalently bind to the specially activated slide surface (“Codalink” GE Healthcare/Amersham Biosciences). We printed these oligos to a test array using a Brown lab spotter and treated the slides to allow the amino linker to covalently attach to the

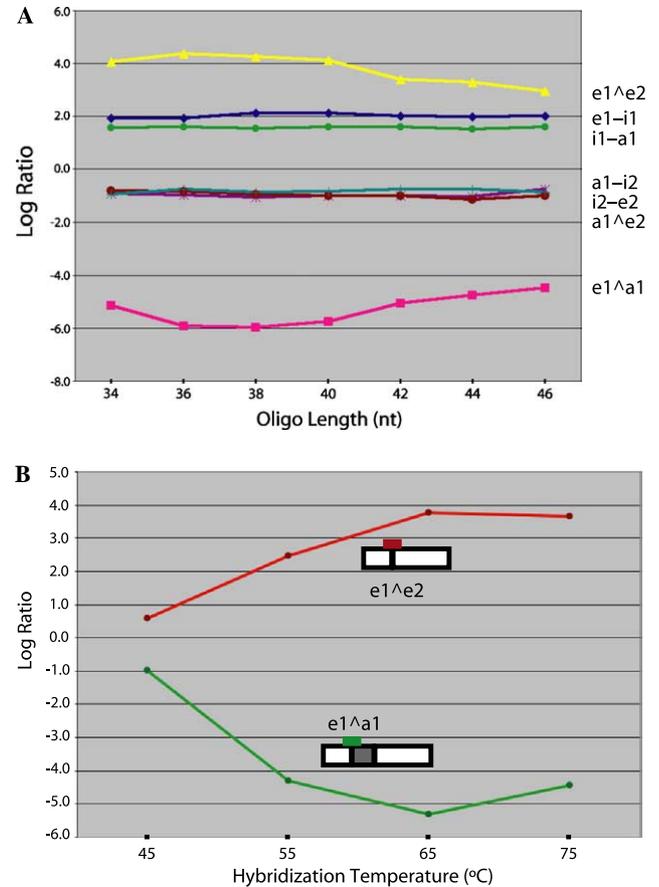


Fig. 5. Relationship of specificity to length and temperature. (A) Data from the spots shown in the scan in Fig. 4 was plotted to measure specificity. The junction probes are shown at right. In theory, probes for a1-i2, i2-e2, and a1^e2 should capture the same amount of signal in both fluor channels, and thus would be centered at 0 if the data was normalized. This accounts for the slightly reddish appearance of these spots in Fig. 4. Note that the best specificity will be when the log<sub>2</sub> ratios of the e1^e2 probe and the e1^a1 probe differ the most. This occurs at lengths of 36–40 under our hybridization conditions. (B) Data from replicate spots for e1^a1 and e1^e2 probes hybridized at different temperatures were averaged and plotted. Note that the best specificity will be when the log<sub>2</sub> ratios of the e1^e2 probe and the e1^a1 probe differ the most. This occurs at temperatures near 65 °C under our hybridization conditions.

slide surface, and then blocked the unreacted groups on the slide with ethanolamine as recommended by the manufacturer.

RNA from yeast carrying the wild type *DYN2* construct was labeled by reverse transcription in the presence of Cy3-dUTP and RNA from a strain carrying the intron 1 branchpoint mutant *dyn2* construct was labeled with Cy5-dUTP. The labeled cDNAs were mixed and hybridized together to arrays overnight at 45, 55, 65, and 75 °C. After washing, arrays were scanned using an Axon 4000A scanner and data was extracted using GenePix software from Axon. Separate arrays were scanned at laser powers determined empirically to give similar intensity value distributions across the arrays for the two dyes and the data was not normalized. Fig. 4C shows false color images of the spot intensities for each series of junction probes hybridized

at 65 °C. Green indicates capture of RNA from the wild type strain, whereas red indicates capture of RNA from the mutant strain. When a probe captures target from both samples, it appears yellow.

### 3.1.3. Specificity for spliced product is temperature and probe length specific

Specificity and sensitivity can be observed to vary with length for several of these probe series. For example, the e1<sup>e</sup>2 probe, which should only capture target from the Cy3-labeled wild type (green) sample, becomes yellow as the probe is made longer, a sign that Cy5-labeled sample from the mutant is annealing (Fig. 4C, compare 34 nt to 48 nt in the first row of spots). The e1<sup>e</sup>3 probe (second row) is red over the entire length series, indicating that the shortest

probe is both efficient and specific, and that increased length is not accompanied by a loss of specificity. The third row shows the e1–i1 junction, which should be found mostly in the mutant (transcript B in Fig. 6), and thus should appear red. The shorter members of the probe set capture little signal, however the longer members can detect the transcript specifically in the mutant. Thus sensitivity of this probe increases with length without a loss of specificity.

Estimating the efficacy of probes by inspecting false color images is not quantitative. We compared intensity measurements and calculated log<sub>2</sub> ratios for the experiment above as. First, we compared intensity (as a measure of sensitivity) and found that as expected larger oligos capture more target at any temperature than do smaller oligos, and that increasing the temperature of hybridization reduces

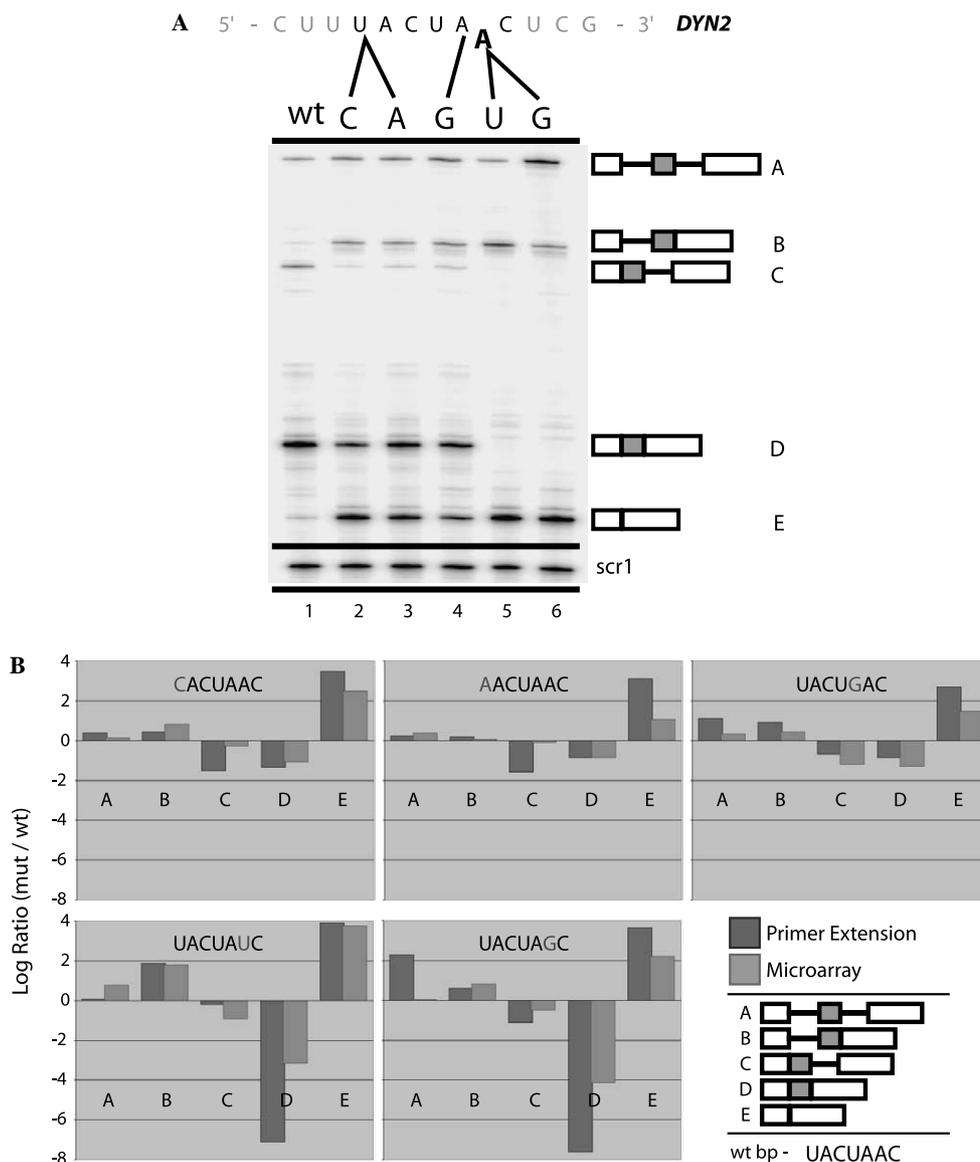


Fig. 6. Comparison of primer extension and microarray in detection of different amounts of exon skipping in *DYN2*. (A) Primer extension with an oligo complementary to exon 3. At the top is the sequence of *DYN2* intron 1 near the branchpoint UACUAAC sequence and the mutations tested. Log<sub>2</sub> ratios were calculated by normalizing each lane to scr1 signal and then determining the ratio of the indicated band (A–E at right) in a mutant lane to the corresponding wild type band. (B) Calculated log<sub>2</sub> ratios from the primer extension (dark gray bars) were plotted next to the log<sub>2</sub> ratios from the microarray (light gray bars) for the mutants indicated (altered base in gray at top of each plot) and for the transcripts indicated (A–E) at bottom.

target capture (data not shown). To determine the specificity of this hybridization, we calculated log ratios for the probes and have summarized the data in Fig. 5. It was important to ask whether longer oligos displayed more crosstalk than shorter probes designed to the same junction. As suggested in Fig. 4C and now shown in Fig. 5A, the specificity of some probes was not affected by length. Conversely, the log ratios for other probes move towards zero at longer oligo lengths, indicating a loss of specificity, even at 65°C. Each oligo is going to have sequence-dependent properties, however these results indicate that a probe length of approximately 38–40 nucleotides provides a reasonable trade-off between signal intensity and specificity.

We also tested probe specificity at different hybridization temperatures (Fig. 5B). We expected that higher temperatures would increase the specificity of target capture, and the results support this hypothesis. At 45 and 55°C,  $\log_2$  ratios from the e1<sup>^</sup>a1 probe (for which target is present only in the Cy3-labeled sample) and that of the e1<sup>^</sup>e2 probes (for which target is present only in the Cy5-labeled sample) are near zero, indicating weak specificity. As hybridization temperature increases, specificity improves and is best at about 65°C (Fig. 5B). Increasing the hybridization temperature to 75°C did not greatly increase the specificity, but did result in a loss of average signal intensity (data not shown), suggesting that sensitivity will be lost for genes with low expression levels. These experiments demonstrate that the basic expectations that temperature and probe length present trade offs between sensitivity and specificity, and that this may be especially challenging for junction oligos due to half-junction crosstalk.

#### 3.1.4. Detection and measurement of subtle changes in alternative splicing

Mutations in the branchpoint of the first intron in the *DYN2* gene cause exon 2 to be skipped [40]. Yeast strains transformed with *DYN2-CUP1* fusion plasmids and carrying the indicated mutations were grown to early log phase, and RNA was isolated. Quantitative reverse transcription with a labeled primer complementary to the e2 exon revealed the different amounts of each of the *DYN2* splicing products, precursors and intermediates (Fig. 6A). The amounts of each were determined by a phosphorimager and normalized to the SRP RNA *scr1* as an internal control. Consistent with previous observations [40], different mutations in the branchpoint region cause different amounts of skipping (Fig. 6). We extracted values representing the detected amounts of each band for each mutant and the wild type. Band intensities from each mutant were related to the corresponding band from wild type in Fig. 6A to create mutant/wt  $\log_2$  ratios that were comparable to microarray data. Data from primer extension experiments correlated well with microarray results (Fig. 6B). Possibly due to limited dynamic range, the microarrays appeared to be less sensitive for the most dramatic ratios. However, the relative magnitudes were retained, and the degree to which alternative splicing takes place in the different mutants was

reflected in the array results. These results suggest that splicing-sensitive microarrays have the potential to detect and measure relatively small changes in levels of alternative splicing in natural samples.

### 3.2. Finding splicing differences between human cancer cell lines

#### 3.2.1. Array design for application to human alternative splicing

We designed a microarray to discover and profile changes in alternative splicing for 64 human genes that were either (1) well expressed in cultured human cells or (2) known to be involved in cancer. To find well-expressed genes, we measured expression in four cultured human cell lines using Affymetrix HG-U133A microarrays. Next, we reconciled this list with a list of human genes with reliable EST or mRNA evidence for alternative splicing [23]. We also identified genes involved in cancer for which changes in alternative splicing have been documented. The list of 64 genes, their functions, the number of regions, and modes of splicing are shown in Table 1.

We designed oligos to cover constitutive exons and alternative exons, restricting the length of these to 40 nucleotides (nt). To learn more about optimizing junction oligos, we tested three methods for selecting a probe sequence. Using the *DYN2* results, we made 40-mers centered (C) on the junction with 20 nt from each exon. In a second strategy, we allowed a window of 40 nt to slide (S) across the target junction to minimize the difference in  $T_m$  of the half-junctions, as calculated using a nearest-neighbor model [38]. A third strategy was to allow the probe sequence to grow from the junction, adding to whichever side had the lower  $T_m$  (wiggle, W) without overshooting a target total  $T_m$  of 91°C. The results of this approach for the alternative exon near the 3' end of the *ITGA6* gene is shown in Fig. 7A. In some cases two algorithms picked the same probe, in which case the probe was named for the simplest algorithm that picked it. An important sanity check on the probe picking process is to use an independent method of confirmation that a probe matches back to the gene or transcript to which it was designed. In the case of the *ITGA6* probes on this array, we could use BLAT to locate these probe sequences in the UCSC Human Genome Browser, alongside the EST/mRNA evidence that demonstrates the alternative splicing of the penultimate *ITGA6* exon (Fig. 7B).

Unlike the *DYN2* experiment above, our primary goal here is to test our ability to detect and measure unknown changes in human alternative splicing. To do this with the greatest chance of success, we must normalize the raw array data using many RNAs that are not expected to change between samples. Thus array features must be designed for control purposes to capture these signals as well. We have used two classes of such control sequences: (1) a set of “stoic” genes [41,42], whose expression in mammalian cells and tissues changes the least across many types of

Table 1  
Summary of genes probed on human splicing test microarray

Gene Names	Description	Number of regions <sup>a</sup>	Splicing modes <sup>a</sup>
ACPI	Phosphotyrosyl protein phosphatase	1	Complex
ADIR	ATP-dependent interferon response protein 1	1	Complex
APLP2, TOR3A	Amyloid $\beta$ (A4) precursor-like protein 2	2	Complex; single cassette
BCL2L1	BCL2-like 1, (bcl-x)	1	alt 5
CACNA1G	Calcium channel, voltage-dependent, $\alpha$ 1G	3	2 single cassettes; double cassette
CASP3	Caspase 3 preproprotein	1	Single cassette
CD44	Cell adhesion molecule	1	Complex
CKLF1	Transmembrane proteolipid (C32)	1	Double cassette
CRYZ	Crystallin, zeta (quinone reductase)	1	Single cassette
CSDA	Cold shock domain protein A	1	Single cassette
DDR1	Tyrosine protein kinase (CAK)	1	Complex
DKFZP586A011, LETMD1	Cervical cancer 1 protooncogene protein p40	2	Single cassette; complex
DNM1L	Dynamin 1-like protein	1	Double cassette
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 $\beta$	2	Single cassette; double cassette
DTYMK	Deoxythymidylate kinase (thymidylate kinase)	1	Single cassette
EEF1D	Translation elongation factor 1 $\delta$	1	Single cassette
ERP28	Endoplasmic reticulum luminal protein	1	Single cassette
ESR1	Estrogen receptor 1	1	Multi-cassette
ESR2	Estrogen receptor2	1	Double cassette
FAS, TNFRSF6	Apoptosis (APO-1) antigen 1	1	2 double cassettes
FASL, TNFSF6, FASLG	Apoptosis (APO-1) antigen ligand 1	1	Single cassette
FLJ10482, RBM23	Hypothetical protein	1	Single cassette
GGCX	$\gamma$ -Glutamyl carboxylase	1	Single cassette
HNRNPAB	hnRNPA, B proteins	1	Single cassette
HNRNPD	Heterogeneous nuclear ribonucleoprotein D	1	Single cassette
HNRPC	Heterogeneous nuclear ribonucleoprotein C	1	alt 5
HRMT1L1	hnRNP methyltransferase-like	1	Single cassette
HT007	Hypothalamus protein HT007	1	Double cassette
HTATIP	Tat interactive protein (TIP60)	1	Single cassette
IFI16	Interferon, $\gamma$ -inducible protein 16	1	Double cassette
ITGA6	Integrin $\alpha$ 6 precursor	1	Single cassette
KARS	Mitochondrial lysyl-tRNA synthetase	1	Single cassette
LARD, TNFRSF25	Tumor necrosis factor receptor family	1	Complex
MAP3K7	Mitogen-activated protein 3(kinase) 7	1	Single cassette
MAP4K4	Mitogen-activated protein 4(kinase)	1	Double cassette
MAPT	Microtubule-associated protein Tau	4	Double cassette; 3 single cassettes
MYL6	Myosin light chain 6	1	Single cassette
MYLK	Myosin light chain kinase	3	3 single cassettes
NASP	Nuclear autoantigenic sperm protein	1	Single cassette
PSEN1	Presenilin proteins	1	Single cassette
PTB1	Pyrimidine tract binding protein	1	Single cassette
PTD013, RWDD1	Function unknown	1	Double cassette
RAD1	Exonuclease homolog RAD1	1	Double cassette
RAD51C	RAD51 ( <i>S. cerevisiae</i> ) homolog C	1	Single cassette
RBM9	RNA binding motif protein 9	2	2 single cassettes
RIC1, ARHGAP17	Homolog of rat nadrin	1	Single cassette
SCML1	Sex comb on midleg-like 1	1	Double cassette
SNAP25	Synaptosomal-associated protein	1	Mutually exclusive cassette
SR-BP1	Sigma receptor (SR31747 binding protein 1)	1	Single cassette
SSBP3	Single-stranded DNA binding protein	1	Double cassette
SSH3BP1, ABI1	Spectrin SH domain binding protein	1	Single cassette
STK6	Serine/threonine kinase 6	1	Complex
TARBP1	TAR (HIV) RNA-binding protein 1	1	Single cassette
TCF3	Helix-loop-helix protein HE47 (E2A)	1	Mutually exclusive cassette
TERF1	Telomeric repeat binding factor 1	1	Single cassette
TIMM8B	Translocase of inner mitochondrial membrane	1	Single cassette
TP73	Tumor suppressor p73	2	2 single cassettes
TPD52L1	Tumor protein D52-like 1	1	Double cassette
TPM2	$\beta$ -Tropomyosin	1	Mutually exclusive cassette
VDU1, USP33	pVHL-interacting deubiquitinating enzyme	1	Single cassette
VEGF	Vascular endothelial growth factor	1	Single cassette
VEGFB	Vascular endothelial growth factor B	1	alt 3
WT1	Wilm's tumor protein	1	Single cassette
ZNF207	Zinc finger protein	1	Single cassette

<sup>a</sup> The terms "region" and "mode" are defined in the text.

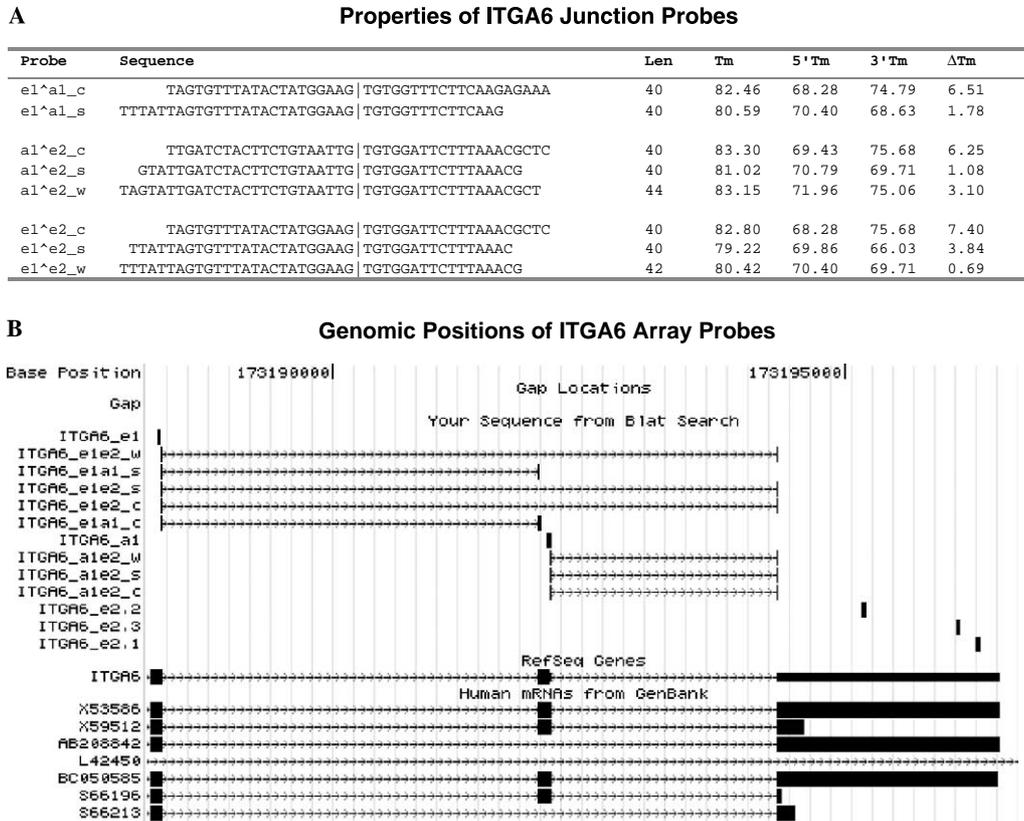


Fig. 7. Properties and genomic positions of splicing-sensitive array probes for ITGA6. (A) Sequences of the probes designed for ITGA6 using three methods, center, slide, and wiggle (indicated by last letter of probe name, see text), and the calculated  $T_m$ , half- $T_m$ s and the difference between the half- $T_m$ s ( $\Delta T_m$ ). (B) Display of the array probe locations relative to observed spliced mRNAs aligned to the genome using UCSC Genome Browser. Probe names, gene name, and cDNA accession numbers are at left, genome position is across the top. Note that e1^a1 has no wiggle probe because the slide algorithm picked the identical oligonucleotide.

experiments and (2) a set of synthetic RNAs whose sequence is not in the human genome and which are added in to the human RNA sample in known amounts prior to the labeling.

Once we decided on the target gene list, picked constitutive and alternative probes for those genes, picked the probes for the stoic genes and the spike-in controls, we ordered the oligonucleotides. Oligos were ordered with 5' amino-linkers from Sigma-GeneSys at 100  $\mu$ M in water in 96 well plates. These were diluted and rearrayed into 384-well plates and printed onto “Codelink” slides (from GE Healthcare/Amersham Biosciences) or “GAPS” slides from Corning. The resulting test array had  $\sim$ 9600 spots with each of 2000 splicing-relevant oligos printed four times each at different locations on the array.

### 3.2.2. Capture and normalization of array data

Total RNA from cancer cell lines was extracted and reverse-transcribed using a mix of oligo(dT) and random primers. Oligo(dT) primers ensure selection of full-length mRNA transcripts, whereas random hexanucleotide primers minimize the 3' bias inherent in reverse transcription primed only with oligo(dT) primers. Arrays were hybridized using cDNA from as little as 20  $\mu$ g of total human cell RNA, or 1  $\mu$ g of poly(A) RNA. Arrays were scanned on an

Axon Scanner (4000A or 4000B) at laser power settings adjusted so that the overall machine counts in the two channels are about the same. Genepix software generates tiff image files in two emission channels during the simultaneous scans for the Cy3 and Cy5 fluors. Data are extracted using GenePix software (Axon).

With the GenePix software, the user applies and aligns a grid layout to the spots, makes background determinations and captures data from the image file. The user can generate flags that identify bad spots or other damaged sections of the array to prevent them from distorting the analysis. Stringent criteria are consistently applied to each array for identification of poor spot quality. Spots are automatically flagged for exclusion from analysis if they have low composite intensities, have a low percentage of reporting pixels, or there is high variation between a feature's mean and median pixel intensity values in either channel, or if there is a high percentage of saturated pixel values in either channel. Additionally, the user scans the image for other array or spot defects that are not caught by the automatic flagging. Thus, there can be a significant user-specific component to the data flagging, and care in training users to be disciplined and principled is essential for uniform analysis.

GenePix then generates a results file (called a .gpr file) that is archived and further analyzed in R using Bioconductor

packages [43]. Briefly, median intensity and background values (local background) for each channel and for each spot are read into R. Background-subtracted values are used to compute the  $M$  ( $\log_2$  ratio) and  $A$  (composite log intensity) values for each spot. Spots that were flagged in GenePix are given  $M$  values of “NA” to exclude them from further analysis.  $M$  values for each slide are loess regression normalized on  $A$  (to a mean of 0), based on a composite normalization scheme of loess using the set of stoic genes [41,42] to represent genes whose expression is generally not expected to vary between the two co-hybridized samples. This normalization scheme assumes that across these genes, the mean log ratio at any particular  $A$  (composite intensity) value should be 0 in each hybridization. Use of a composite normalization mitigates any extreme  $M$  value correction that either normalization alone might do, especially for intensity dependent artifacts.

Each pair of samples is hybridized to two arrays with dye reversal, and the normalized log ratios ( $M$ ) from replicate, dye-reversed hybridizations are averaged. Quality scores are calculated for each average log ratio value, based on  $A$  value, the percentage of spot replicates reporting a value, and the coefficients of variation of the ratios in the average. These quality scores can be used in downstream analyses (correlation, etc.). Other slide-based quality measures are inspected by the user to aid in decisions about whether to include an array in later analyses. These measures include percentage of the oligos that produced high quality signals (as above), overall range of intensities,  $\log_2$  ratios, and background.

### 3.2.3. Indexes derived from within-gene normalization detect splicing changes

In standard gene expression microarray analysis, once the array data is normalized, the identification of genes whose transcript levels change can begin. For those of us whose interests run to splicing however, the work is not yet done. To be sure, we can now evaluate expression level changes in our experiments by using the constitutive probe signals, however splicing changes may not yet be evident. To extract splicing changes we create “indexes” which are comparisons in the signal between different regions in the gene model [11]. Most often we would compare the signals from the constitutive regions (as a measure of the overall transcript level) to the signals from the alternatively spliced regions of the transcripts. Since we are dealing with changes in the ratio between one sample and another (expressed as the  $\log_2$  of the ratio Sample/Reference), we must create a ratio of ratios between one part of the transcript and another (expressed as the  $\log_2$  of the ratio of the ratio Sample alternative features/Reference alternative features) by the ratio (Sample constitutive features/Reference constitutive features). In other words we can subtract the  $\log_2$  ratio value of the constitutive features from the  $\log_2$  ratio of the alternative features, to bring all the alternative features into line with each other for comparison of their signals to those of the overall transcript pool. We have done this “within-

gene normalization” with data for two genes, ITGA6 encoding the integrin  $\alpha$  6 precursor, and MYL6, encoding myosin light chain 6. The experiment compares the teratocarcinoma-derived cell line NCCIT (ATCC No. CRL-2073) to the malignant skin melanoma-derived cell line A375 (ATCC No. CRL-1619), using a common arbitrary reference RNA (Table 2).

To illustrate this we show the normalized array data for each of two pairs of slides for each probe we designed for ITGA6 and MYL6. The columns marked F have the Cy5-label on the experimental sample with Cy-3 on the reference sample, and in the columns marked R the dyes have been reversed. The numbers in these four columns are the average  $\log_2$  ratios of the four (or fewer) spots for the indicated array feature (oligo). Comparison of equivalent quality dye-swapped replicates can provide an estimate of the reproducibility of each feature’s performance, since the RNA samples should be the same, with the variation restricted to the original slides and any dye-specific effects, which are usually rare. In the case of several probes, data was not recovered, indicating that the spots for these oligos did not meet quality criteria and were flagged.

To perform within-gene normalization, we averaged the  $\log_2$  ratios for all the constitutive features predicted by the gene model (indicated with C in the Type column). This value appears in a box in the within-gene normalized columns at the right (Table 2). This value represents the  $\log_2$  ratio of the relative amounts of ITGA6 mRNA in NCCIT (or A375) cells as compared to the reference cell line (HEK293 cells in this example). According to this value, both cell lines produce about 8-fold more ITGA6 mRNA than the reference. Since the reference is common to this experiment, we can also infer that A375 cells produce less than 2-fold more ITGA6 mRNA than do NCCIT cells. We then normalize the expression ratios for the alternative (I, include, these are shaded; S, skip) features for ITGA6 by subtracting the average  $\log_2$  ratio of the constitutive features (2.88 for NCCIT cells, 3.21 for A375 cells) from the  $\log_2$  ratio of each alternative feature.

Having normalized the alternative features to overall expression level, we can now compare them to each other. We averaged the normalized include features (except for those with asterisks since these gave unreliable signals) and the skip features, and these are shown in the box at lower right for each gene. According to these numbers, NCCIT cells have more than 8-fold less inclusion (on a per transcript basis) of the penultimate ITGA6 exon than do HEK293 cells, and have perhaps 1.5-fold more skipped ITGA6 mRNA (on a per transcript basis) than do HEK cells. We can use the HEK reference to estimate the skip/include ratio (S/I) by subtracting the average include  $\log_2$  ratio from the average skip  $\log_2$  ratio. This value,  $-3.91$ , suggests a skip to include ratio of just less than 16 in the ITGA6 mRNA pool of NCCIT cells. Further analysis suggests that inclusion is 2-fold more common than skipping in the ITGA6 mRNA pool within A375 cells. The MYL6 gene has nearly identical expression levels in all

Table 2  
Alternative Splicing Array Data

Array feature	Type	NCCIT		A375		Within-gene normalized		
		F	R	F	R	NCCIT	A375	
ITGA6_e1	C	2.97	3.23	3.29	3.80	2.88	3.21	
ITGA6_e2.1	C	3.12	2.61	3.60	3.62			
ITGA6_e2.2	C	2.50	2.36	3.20	2.10			
ITGA6_e2.3	C	3.86	2.39	3.74	2.34			
ITGA6_a1	I	0.71	0.89	4.22	4.45	−2.08	1.13	
ITGA6_a1^e2_c	I	0.56	−0.64	4.19	4.68	−2.92	1.22	
*ITGA6_a1^e2_s	I	—	—	4.56	—	—	1.35	
ITGA6_a1^e2_w	I	−1.09	−1.20	4.20	4.19	−4.03	0.99	
ITGA6_e1^a1_c	I	—	−1.24	4.31	4.75	−4.12	1.32	
*ITGA6_e1^a1_s	I	—	—	4.89	4.40	—	1.44	
ITGA6_e1^e2_c	S	3.20	2.82	3.71	3.41	0.13	0.35	
ITGA6_e1^e2_s	S	4.03	3.56	2.66	2.99	0.92	−0.38	
ITGA6_e1^e2_w	S	3.94	3.44	3.04	4.61	0.81	0.62	
						Avg Include	−3.29	1.17
						Avg Skip	0.62	0.20
						Skip-Incl	3.91	−0.97
MYL6_e1	C	0.42	0.18	0.19	0.04	0.32	0.04	
MYL6_e2	C	0.37	0.39	0.13	0.11			
MYL6.1	C	0.42	0.11	0.21	−0.19			
MYL6.2	C	0.39	0.27	0.10	−0.24			
MYL6.3	C	0.42	0.24	0.15	−0.15			
MYL6_a1	I	−1.41	−1.13	1.02	0.70	−1.59	0.83	
MYL6_a1^e2_c	I	−0.73	−0.68	0.04	−0.37	−1.03	−0.20	
MYL6_a1^e2_s	I	0.33	0.05	0.29	0.25	−0.13	0.23	
MYL6_a1^e2_w	I	−2.94	−2.82	1.42	1.68	−3.20	1.51	
MYL6_e1^a1_c	I	−1.45	−1.00	1.27	0.89	−1.54	1.05	
MYL6_e1^a1_s	I	−1.54	−1.77	1.23	0.91	−1.98	1.03	
MYL6_e1^a1_w	I	−1.16	−0.76	1.10	0.79	−1.28	0.91	
MYL6_e1^e2_c	S	0.61	0.23	0.10	−0.11	0.10	−0.04	
MYL6_e1^e2_s	S	0.73	0.47	−0.23	−0.75	0.28	−0.53	
						Avg Include	−1.54	0.77
						Avg Skip	0.19	−0.29
						Skip-Incl	1.73	−1.06

Array features are named by gene and position (e1 is first constitutive exon, a1 is first alternative exon, and the ^ symbol indicates a junction). Types: C, constitutive; I, included; S, skipped. Dye labeling: F, forward (test with Cy5, reference HEK with Cy3); R, reverse (test with Cy3 and reference with Cy5). The numbers in the large cells in columns at right are the averages of the constitutive features from the combined arrays. These values have been subtracted from the signals derived from include and skip features for “within-gene” normalization. In the box at lower right for each gene, the averaged, normalized skip and include signals have been determined and a skip/include ratio is derived by subtraction in log space. All numbers are log<sub>2</sub> values. Array features marked with \* were not used to create the average include values during within-gene normalization.

three cell lines, however after examining splicing of the cassette exon studied here, it seems clear that inclusion of the exon differs by almost 8-fold between the two cell lines (Table 2).

### 3.2.4. Comparing selected array results to truth as revealed by RT-PCR

To test the robustness of array predictions, we validated some of the array data by semi-quantitative RT-PCR and in some cases by quantitative PCR (qPCR). Fig. 8 shows agarose gels of RT-PCR results for ITGA6 and MYL6 from cell lines NCCIT and A375. Although these RT-PCRs are subject to effects that limit their use in quantifying the numbers of differently spliced isoforms, they agree remarkably with the microarrays. The array suggests that the S/I ratio of ITGA6 should be nearly 16 in NCCIT cells and about 0.5 in A375 cells. By RT-PCR, no exon-included

form of ITGA6 can be observed in NCCIT cells. Slightly more (on a molar basis) skipped mRNA than include mRNA for ITGA6 is detectable in A375 cells. For the MYL6 gene, the array indicates an S/I ratio of nearly 4 for NCCIT cells, and again, no include mRNA can be detected by RT-PCR, whereas something near the expected 2-fold greater inclusion than skipping is observed for A375 cells. An extra band of 210 nucleotides (marked by an asterisk) is observed in mRNA from A375 cells that was not predicted by our gene model. In the public EST data clones exist which apparently arise from the use of an alternative 5' splice site downstream of the expected exon, leading to a larger PCR product.

We obtained clones for included and skipped transcripts of ITGA6 from ATCC and designed PCR primers to separately and specifically amplify each of the two alternatively spliced transcripts. We generated a standard

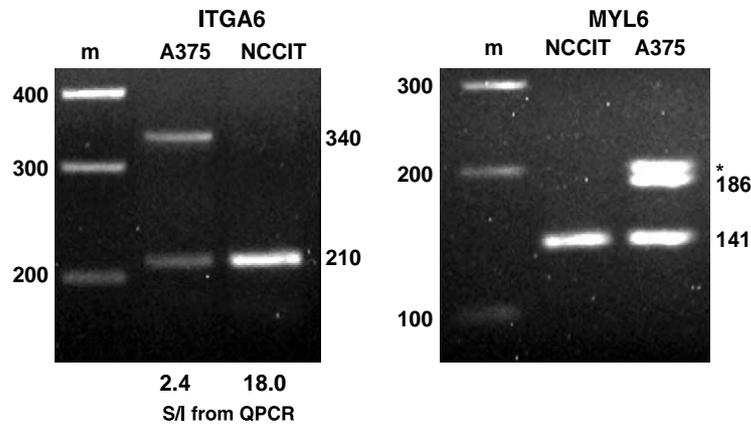


Fig. 8. RT-PCR validation of two alternative splicing differences discovered between NCCIT cells and A375 cells. Standard “semi-quantitative” RT-PCR was carried out using RNA from each cell line and primer pairs complementary to the constitutive exons flanking the alternative exon in each case. The asterisk (\*) marks a ~210 nt MYL6 product in A375 cells that was not in our original gene model. EST evidence in the database suggests this is an alternative form that uses an alternative 5' splice site near the alternative exon. ITGA6 was also studied by qPCR using primer pairs specific for either the included or skipped isoform. Numbers of molecules of each form were determined using dilutions of cDNA clones containing the appropriate isoform as a standard curve. S/I is given below the lane for the indicated cell line and represents the ratio calculated using the absolute number of molecules detected by qPCR.

curve for each isoform using known amounts of target molecules, in this case provided by dilutions of the plasmid DNAs. We performed qPCR using the same set of primers on cDNA templates derived from RNA, to measure the number of each of the alternatively spliced ITGA6 transcripts present in each cell line. These data are shown as a calculated skip/include ration below the gel in Fig. 8. In the case of NCCIT cells the arrays agree with qPCR. In the case of A375 cells, the arrays appear to underestimate the amount of skipped isoform for ITGA6. Given the apparent crosstalk potential at the ITGA6 junctions (Figs. 3 and 7), it seems likely that one or several of the features have reduced specificity, and that averaging these may influence the final predicted ratio.

#### 4. Conclusions

The application of microarrays to splicing is still in a stage of development. The early experiments are promising, and it is very clear that for a large majority of genes and splicing events, microarrays will operate well to detect splicing changes. The numbers of genes and splicing events that can be accessed in a single experiment approaches a hundred to a thousand fold more than can reasonably be accessed using RT-PCR, primer extension, or nuclease protection experiments. This huge increase in reach promises to allow discovery of splicing regulatory networks on a genomic scale, and may lead to new insights into the regulation of this important step in gene expression.

There is room for improvement in the technology that will make it more cost effective for small groups to use to ask focused experimental questions. In particular, design of junction oligos could be improved, especially where splice junctions are difficult to discriminate due to similarity between alternative exons, or other structure near the splice junctions (e.g., ITGA6, Figs. 3 and 7). Furthermore the

methods for statistical analysis of signals derived from splicing sensitive microarrays have not yet been made sophisticated. For example, the within-gene normalization discussed in this article employs simple averaging of signals from each feature that provides a signal, but more elaborate filtering, identifying and increasing the weight of signals from “good” features, or use of other measures of central tendency for normalization could easily be applied. Abundant opportunity exists for analyzing the data with a variety of informatic techniques.

#### Acknowledgments

Funding for the development of splicing-sensitive microarrays in our laboratory has been generously provided by the University of California Cancer Research Coordinating Committee, the Packard Foundation, the W.M. Keck Foundation (through its support of the RNA Center at UCSC), NIH Grant R01 GM 040478 to M. Ares, and R24 GM 070857 to D. Black, M. Ares, and X.-D. Fu. Support for the UCSC Microarray Facility also comes from NHGRI P41 HG02371 to David Haussler. Funding for students in the Hughes Undergraduate Research Lab was generously provided by HHMI.

#### References

- [1] B. Modrek, C. Lee, *Nat. Genet.* 30 (2002) 13–19.
- [2] C.W. Smith, J. Valcarcel, *Trends Biochem. Sci.* 25 (2000) 381–388.
- [3] D.L. Black, *Annu. Rev. Biochem.* 72 (2003) 291–336.
- [4] D.L. Black, P.J. Grabowski, *Prog. Mol. Subcell. Biol.* 31 (2003) 187–216.
- [5] N.A. Faustino, T.A. Cooper, *Genes Dev.* 17 (2003) 419–437.
- [6] P.J. Grabowski, D.L. Black, *Prog. Neurobiol.* 65 (2001) 289–308.
- [7] R.P. Auburn, D.P. Kreil, L.A. Meadows, B. Fischer, S.S. Matilla, S. Russell, *Trends Biotechnol.* 23 (2005) 374–379.
- [8] L. Qin, L. Rueda, A. Ali, A. Ngom, *Appl. Bioinformatics* 4 (2005) 1–11.
- [9] G. Sherlock, C.A. Ball, *Mol. Biotechnol.* 30 (2005) 239–252.

- [10] T. Burckin, R. Nagel, Y. Mandel-Gutfreund, L. Shiue, T.A. Clark, J.L. Chong, T.H. Chang, S. Squazzo, G. Hartzog, M. Ares Jr., *Nat. Struct. Mol. Biol.* 12 (2005) 175–182.
- [11] T.A. Clark, C.W. Sugnet, M. Ares Jr., *Science* 296 (2002) 907–910.
- [12] J. Ule, A. Ule, J. Spencer, A. Williams, J.S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B.R. Zeeberg, D. Kane, J.N. Weinstein, J. Blume, R.B. Darnell, *Nat. Genet.* (2005).
- [13] H. Wang, E. Hubbell, J.S. Hu, G. Mei, M. Cline, G. Lu, T. Clark, M.A. Siani-Rose, M. Ares, D.C. Kulp, D. Haussler, *Bioinformatics* 19 (Suppl. 1) (2003) i315–i322.
- [14] Q. Pan, O. Shai, C. Misquitta, W. Zhang, A.L. Saltzman, N. Mohammad, T. Babak, H. Siu, T.R. Hughes, Q.D. Morris, B.J. Frey, B.J. Blencowe, *Mol. Cell* 16 (2004) 929–941.
- [15] M. Blanchette, R.E. Green, S.E. Brenner, D.C. Rio, *Genes Dev.* 19 (2005) 1306–1314.
- [16] J. Castle, P. Garrett-Engele, C.D. Armour, S.J. Duenwald, P.M. Loerch, M.R. Meyer, E.E. Schadt, R. Stoughton, M.L. Parrish, D.D. Shoemaker, J.M. Johnson, *Genome Biol.* 4 (2003) R66.
- [17] J.M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P.M. Loerch, C.D. Armour, R. Santos, E.E. Schadt, R. Stoughton, D.D. Shoemaker, *Science* 302 (2003) 2141–2144.
- [18] K. Le, K. Mitsouras, M. Roy, Q. Wang, Q. Xu, S.F. Nelson, C. Lee, *Nucleic Acids Res.* 32 (2004) e180.
- [19] J.M. Yeakley, J.B. Fan, D. Doucet, L. Luo, E. Wickham, Z. Ye, M.S. Chee, X.D. Fu, *Nat. Biotechnol.* 20 (2002) 353–358.
- [20] B. Modrek, A. Resch, C. Grasso, C. Lee, *Nucleic Acids Res.* 29 (2001) 2850–2859.
- [21] T.A. Thanaraj, S. Stamm, F. Clark, J.J. Riethoven, V. Le Texier, J. Muilu, *Nucleic Acids Res.* 32, Database issue (2004) D64–D69.
- [22] C. Grasso, B. Modrek, Y. Xing, C. Lee, *Pac. Symp. Biocomput.* (2004) 29–41.
- [23] C.W. Sugnet, W.J. Kent, M. Ares Jr., D. Haussler, *Pac. Symp. Biocomput.* (2004) 66–77.
- [24] R. Sorek, G. Ast, *Genome Res.* 13 (2003) 1631–1637.
- [25] N. Kim, S. Shin, S. Lee, *Genome Res.* 15 (2005) 566–576.
- [26] K. Malde, E. Coward, I. Jonassen, *Bioinformatics* 21 (2005) 1371–1375.
- [27] Y. Xing, A. Resch, C. Lee, *Genome Res.* 14 (2004) 426–441.
- [28] S. Heber, M. Alekseyev, S.H. Sze, H. Tang, P.A. Pevzner, *Bioinformatics* 18 (Suppl. 1) (2002) S181–S188.
- [29] C. Lee, C. Grasso, M.F. Sharlow, *Bioinformatics* 18 (2002) 452–464.
- [30] C.A. Davis, L. Grate, M. Spingola, M. Ares Jr., *Nucleic Acids Res.* 28 (2000) 1700–1706.
- [31] M. Spingola, L. Grate, D. Haussler, M. Ares Jr., *RNA* 5 (1999) 221–234.
- [32] D.D. Shoemaker, E.E. Schadt, C.D. Armour, Y.D. He, P. Garrett-Engele, P.D. McDonagh, P.M. Loerch, A. Leonardson, P.Y. Lum, G. Cavet, L.F. Wu, S.J. Altschuler, S. Edwards, J. King, J.S. Tsang, G. Schimmack, J.M. Schelter, J. Koch, M. Ziman, M.J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M.R. Meyer, M. Mao, J. Burchard, M.J. Kidd, H. Dai, J.W. Phillips, P.S. Linsley, R. Stoughton, S. Scherer, M.S. Boguski, *Nature* 409 (2001) 922–927.
- [33] E.K. Nordberg, *Bioinformatics* 21 (2005) 1365–1370.
- [34] A. Relogio, C. Schwager, A. Richter, W. Ansorge, J. Valcarcel, *Nucleic Acids Res.* 30 (2002) e51.
- [35] J.M. Rouillard, C.J. Herbert, M. Zuker, *Bioinformatics* 18 (2002) 486–487.
- [36] E. Southern, K. Mir, M. Shchepinov, *Nat. Genet.* 21 (1999) 5–9.
- [37] R.J. Britten, E.H. Davidson, *Proc. Natl. Acad. Sci. USA* 73 (1976) 415–419.
- [38] J. SantaLucia Jr, H.T. Allawi, P.A. Seneviratne, *Biochemistry* 35 (1996) 3555–3562.
- [39] K.J. Breslauer, R. Frank, H. Blocker, L.A. Marky, *Proc. Natl. Acad. Sci. USA* 83 (1986) 3746–3750.
- [40] K.J. Howe, C.M. Kane, M. Ares Jr, *RNA* 9 (2003) 993–1006.
- [41] L.L. Hsiao, F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee, K.E. Clark, P. Haverty, Z. Weng, G.L. Mutter, M.P. Frosch, M.E. Macdonald, E.L. Milford, C.P. Crum, R. Bueno, R.E. Pratt, M. Mahadevappa, J.A. Warrington, G. Stephanopoulos, S.R. Gullans, *Physiol. Genomics* 7 (2001) 97–104.
- [42] J.A. Warrington, A. Nair, M. Mahadevappa, M. Tsyganskaya, *Physiol. Genomics* 2 (2000) 143–147.
- [43] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, J. Zhang, *Genome Biol.* 5 (2004) R80.