Minireview

Analysis of alternative splicing with microarrays: successes and challenges

Christopher Lee and Meenakshi Roy

Address: Molecular Biology Institute, Center for Genomics and Proteomics, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095-1570, USA.

Correspondence: Christopher Lee. E-mail: leec@mbi.ucla.edu

Published: 21 June 2004

Genome ***Biology*** 2004, **5:**231

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/7/231

Abstract

Recently, DNA microarrays have emerged as potentially powerful tools for analyzing alternative splicing. We briefly review the latest results in this field and highlight the current challenges that they have revealed.

The field of genomics is sometimes accused of being largely a numbers game - increasing our knowledge quantitatively without adding qualitatively to our conceptual understanding. But sometimes big numbers change our mental models. One area in which genomic data appear to be causing just such a shift is the field of alternative splicing. The 'one gene, one product' dogma of molecular biology is yielding in the face of large amounts of human genome data to 'most genes have multiple products', with important implications throughout biology [1-6]. Recently, several large-scale studies [7-9] have shown that alternative splicing can be analyzed in a high-throughput manner using DNA-microarray methods, an approach that is likely to be useful for understanding the role of alternative splicing in many areas of biology.

Bioinformatic analyses of expressed sequence tag (EST) data were the first to herald the alternative-splicing revolution. A number of studies by different groups all reported finding alternative splice forms in a surprisingly large fraction of human genes, ranging from 40% to 60% [10-15]. These studies have identified more than 30,000 alternative splice forms in human, effectively doubling the number of human gene products relative to the estimated 32,000 human genes. But EST data clearly do not tell the whole story. Even assuming that a wide variety of potential problems are carefully filtered out (for example, genomic contamination and incomplete mRNA processing; see [16]), the very nature of the EST data leaves many questions unanswered. Individual ESTs might represent rare splice forms (or even errors made by the splicing machinery) that do not constitute a significant fraction of the gene's transcripts in living cells. EST sequencing also has some bias and does not evenly cover every part of every gene. One basic constraint on the discovery of alternative splice forms is that there simply aren't enough EST data to give good coverage of most gene regions in anything approaching a representative list of tissues. Even when alternative splice forms are found, information about their tissue-specific regulation is often poor or unavailable.

The use of DNA microarray technology is very attractive for large-scale studies of alternative splicing. By measuring the relative amounts of distinct splice forms in a variety of tissues, microarrays could both test whether a novel splice form really constitutes an important fraction of the gene's transcripts in at least some cell types, and reveal its patterns of regulation across a large number of different tissues. This is very much needed.

Taking full advantage of microarray technology to analyze alternative splicing poses many challenges for current methodologies. Traditional microarrays are designed to measure the total level of expression of a gene, without attempting to distinguish between different splice forms (for a review, see [17]). For example, probe designs and labeling

protocols used for microarray experiments tend to be biased towards the 3′ end of the gene [18]. As each gene is assumed to be expressed as a unit, this is not considered to be a problem. By contrast, for alternative splicing it is important to have probes throughout all regions of the gene - everywhere that splicing might occur. And given that changes in splicing can be subtle (for example, shifting a single splice donor site by 20 nucleotides or fewer), standard probes designed to match an individual exon are inadequate: probes also need to be designed to match each specific exon-exon junction that might be spliced together by an alternative splicing event.

Alternative splicing also poses new challenges for microarray data analysis. The overall expression level of a gene can be represented by a single number and can be measured with reasonable accuracy by averaging the signals of many probes for the gene [19]. Individual probes that diverge significantly from the average profile are generally considered to be outliers and are excluded from the analysis [20]. But such 'inconsistent' results (in which a subset of probes show a large change in signal that is not seen in other probes for the gene) are exactly what alternative splicing will cause. Thus, our challenge is to demonstrate that the probes considered by standard expression-data analysis to be 'noise' are actually reproducible signals, indicative of different patterns of regulation of multiple splice forms.

Despite these challenges, there is now broadly reproducible evidence that alternative splicing can be detected using microarrays. For example, Hu *et al.* [18] used standard Affymetrix array designs to search for evidence of alternative splicing in 1,600 rat genes, performing hybridizations with 10 normal tissue samples. They found that 268 genes (17%) showed signs of alternative splicing, and validation by reverse-transcriptase PCR (RT-PCR) indicated that about half of these represented genuine alternative-splicing events. This work [18] clearly demonstrates that microarrays can detect alternative splicing, but many types of alternative splicing have probably been missed in this study because of technical limitations such as 3′ labeling bias and the absence of probes designed to detect splice junctions.

Additional studies have focused on individual genes with known alternative-splicing patterns, in order to demonstrate that the technology is sufficiently sensitive and reliable. Clark *et al.* [21] used a cDNA spotted array to demonstrate successful detection of experimentally induced intron retention in a number of genes containing introns. Yeakley *et al.* [22] detected alternative splicing in six human genes using a fiber-optic microarray platform. Wang *et al.* [23] performed a quantitative analysis of distinct splice forms of two human genes (*CD44* and *TPM2*) using an Affymetrix microarray platform. Castle *et al.* [24] reported studies of two human genes (*RB1* and *ANXA7*), examining in great detail the experimental factors that determine the response of probes as a function of their distance from an exon junction, their position with the gene, and so on. They also described a novel unbiased protocol for amplification and labeling of full-length RNAs, combining random-primed first-strand and second-strand synthesis steps with an amplification strategy that uses both PCR and *in vitro* transcription. The method is reported to sample the entire transcript and thus prevent the usual bias towards the 3′ end; detection of alternative splicing in the middle or the 5′ end of a gene is thus facilitated. Finally, Neves *et al.* [25] used a microarray to interrogate different exon variants of three alternatively spliced cassette exons in the *Drosophila DSCAM* gene.

Recently, two large-scale microarray studies of alternative splicing have been published [8,9]. Johnson *et al.* [8] designed 36-mer probes complementary to every consecutive exon-exon junction in more than 10,000 multi-exon genes and used an array of the probes to sample expression of splice forms in 52 human tissues, seeking evidence of exon-skipping events. When individual exon-junction probes were significantly downregulated relative to the other probes for the gene, those with statistical confidence above a threshold level were reported as alternative-splicing predictions. Out of a random sample of 153 exon-skipping events predicted by the microarray analysis, 73 were successfully validated by RT-PCR and sequencing (a 48% validation rate). This initial study has made a very substantial contribution to the discovery of alternative splice forms. For genes in which alternative forms had not previously been reported by EST studies, Johnson *et al.* [8] reported that about half showed microarray evidence of exon skipping. Taking into account the rate of validation by RT-PCR, this means that alternative splicing has been discovered in nearly 800 genes that were not previously known to be alternatively spliced [8].

Combining these novel discoveries with alternative splicing results previously identified from ESTs and mRNA sequences, Johnson *et al.* [8] arrived at an estimate that 74% of human multi-exon genes show experimental evidence of alternative splicing. It should be emphasized that this estimate is not an independent validation of EST-based estimates of the extent of alternative splicing, because it includes those EST results in the total estimate, and the EST data actually represent the largest component of this estimate. Indeed, among genes for which no alternative splicing was previously identified by ESTs, genuine alternative splicing was estimated to be found in only about 20% of the genes. This does not contradict the 74% figure of Johnson *et al.* [8]: genes that have failed to show alternative splice forms in previous large-scale mRNA and EST datasets should indeed be less likely than the 'average gene' to have alternative splicing. So what light do the data of Johnson *et al.* [8] shed on the previous results from EST analysis? They provide direct evidence of two problems with EST data. First, the likelihood of observing ESTs for alternative splice forms in a gene correlates with increasing numbers of ESTs

for that gene; it is highest for highly expressed genes and virtually nil for low-abundance genes. The latter clearly present an opportunity for microarray-based detection to make a big contribution. Second, ESTs are two-fold less likely to detect alternative splice events in the middle of a transcript than at its 5′ and 3′ ends. These problems are not surprising.

Researchers using Affymetrix microarrays have also reported large-scale microarray studies of alternative splicing on chromosomes 21 and 22 [7,9]. Using probes spaced approximately every 35 base-pairs (bp) along these chromosomes, they surveyed transcripts from 11 different human cell lines, identifying both novel regions of transcription and apparent changes in exon-inclusion patterns between different cell types. In a recent analysis of these data [9], they reported that the vast majority of known genes on chromosomes 21 and 22 had multiple isoform profiles (a profile was defined as a substantially different combination of probes that give a positive hybridization signal in the cell lines surveyed). Indeed, only 12-21% of genes appeared to have a single isoform profile in all cells, implying that 80% or more of human genes may be alternatively spliced. As this result is based entirely on the microarray data, it does constitute an independent test of the high level of alternative splicing observed in the EST data. RT-PCR of the novel transcript fragments detected by this microarray study validated 63% of those tested, lending general support to the data. It should be noted, however, that these validation tests concentrated on regions of novel transcript fragments distant from known genes; these probably overlap poorly with the novel alternative-splicing results, which were obtained from known genes. It may be reasonable to expect that novel exons in known genes are likely to be validated at the same or higher rate than the newly detected fragments distant from known genes. This study [8,9] did not focus on alternative splicing, however, and did not present RT-PCR validation data specifically for the putative alternative-splicing predictions.

These large-scale studies illustrate nicely the powerful results that microarrays can bring to the study of alternative splicing, but they also show the challenges of the task. It is significant that both studies [7-9] addressed only one kind of alternative splicing: monitoring individual exon inclusion as an on-off event. The Johnson *et al.* study [8] was explicitly designed to detect exon-skipping events, in which a known exon is selectively skipped in one or more tissues. If a novel exon were selectively included (inserted) in certain tissues, however, this array design would probably miss it. The many other types of alternative splicing (alternative 5′ and/or 3′ splice-site usage, mutually exclusive exons, alternative initiation, alternative termination, and so on) were also not considered in this design [8]. Generally speaking, the type of array design used by Johnson *et al.* [8] depends on knowing a complete list of exons and splice forms to look for. Novel exon forms or splices (those not explicitly included in the array design) are by definition mostly invisible. Systematically adding more probes by scanning through the genomic sequence (as in the Affymetrix design [7,9]) can help to identify novel exons.

Detection of novel splice forms also poses a combinatorial problem. Many alternative-splicing events involve only a subtle shift in splice patterns that cannot be tracked well by exon probes (probes designed to match a specific exon). For example, consider a form of an mRNA, missing one exon, that ordinarily constitutes only 1% of a gene's transcripts. If this 'exon-skip' form is upregulated 10-fold in one tissue, exon probes will show at most a 10% change in this tissue, a very small shift that is hard to detect reliably. By contrast, a splice probe (a probe designed to match a specific exon-exon junction in the spliced transcript) that detects only the exon-skip form will show a 1,000% increase. Designing probes for splices between all possible pairs of exons in a gene is impractical; thus, bioinformatic analysis will be required to pick good candidates, which is by no means a trivial problem. Although in principle the dense tiling of probes used on the Affymetrix chip [7] can detect a wider range of alternative splicing types than just exon skipping, it is unclear whether the data will be readily interpretable. It will take quite a bit more experience with these types of arrays to show convincingly that they can identify a specific alternative-splicing event and distinguish it reliably from other possibilities.

And this brings us to the real challenge of the splicing array experiments: data analysis and biological interpretation. These data pose an interesting mix of problems: superficially, the array data appear to show quantitative changes (some expression levels go up while others go down), but as we and others have shown, they actually signal qualitative changes (the existence of two or more distinct splice forms rather than a single category of transcript), which in turn have a deeper structure of relationships best represented using graph theory (that is, full-length isoforms are the set of possible paths through the directed graph in which exons are nodes and splice forms are edges) [26,27]. These are three very different views of the problem that are not ordinarily combined, but for alternative splicing the connections between them can be ignored only at the risk of forgetting one or another critical aspect of the data. The reliable, automatic interpretation of splicing array data (at the very least, to identify specific splice events and isoform sequences) is just one immediate example of this challenge.

The 'one gene, one product' dogma has been built in to the fundamental assumptions of many databases and analysis methods for one compelling reason: it's simple. Are we ready for the complexities of 'one gene, many products' and for all the data required to track these many forms? Not quite. The Human Genome Project's success and its value to many researchers has come from a shared infrastructure of online community databases and resources, which have been centrally supported. Alternative splicing, by contrast, has never had the equivalent of a 'human transcriptome project' and

still lacks much of this community infrastructure. More than anything else, alternative splicing requires a community annotation infrastructure: to share data about known forms; to design experiments for detecting novel forms and share the resulting data; and to annotate the functional significance of known forms as a community effort, with research done independently throughout the community, but shared and integrated centrally.

References

1. Jiang ZH, Wu JY: **Alternative splicing and programmed cell death.** *Proc Soc Exp Biol Med* 1999, **220:**64-72.
2. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL: ***Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity.** *Cell* 2000, **101:**671-684.
3. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: **Increase of functional diversity by alternative splicing.** *Trends Genet* 2003, **19:**124-128.
4. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.** *Proc Natl Acad Sci USA* 2003, **100:**189-192.
5. Modrek B, Lee C: **Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation/loss.** *Nat Genet* 2003, **34:**177-180.
6. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C: **Assessing the impact of alternative splicing on domain interactions in the human proteome.** *J Proteome Res* 2004, **3:**76-83.
7. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR: **Large-scale transcriptional activity in chromosomes 21 and 22.** *Science* 2002, **296:**916-919.
8. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302:**2141-2144.
9. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, *et al.*: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14:**331-342.
10. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9:**1288-1293.
11. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** *FEBS Lett* 2000, **474:**83-86.
12. Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick JS: **ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome.** *Nat Genet* 2000, **24:**340-341.
13. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
14. Kan Z, Rouchka EC, Gish WR, States DJ: **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.** *Genome Res* 2001, **11:**889-900.
15. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucleic Acids Res* 2001, **29:**2850-2859.
16. Modrek B, Lee C: **A genomic view of alternative splicing.** *Nat Genet* 2002, **30:**13-19.
17. Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, **1:**951-960.
18. Hu GK, Madore SJ, Moldover B, Jatkoe T, Balaban D, Thomas J, Wang Y: **Predicting splice variant from DNA chip expression data.** *Genome Res* 2001, **11:**1237-1245.
19. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21:**20-24.
20. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98:**31-36.
21. Clark TA, Sugnet CW, Ares MJ: **Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays.** *Science* 2002, **296:**907-910.
22. Yeakley JM, Fan JB, Doucet D, Luo L, Wickham E, Ye Z, Chee MS, Fu XD: **Profiling alternative splicing on fiber-optic arrays.** *Nat Biotechnol* 2002, **20:**353-358.
23. Wang H, Hubbell E, Hu JS, Mei G, Cline M, Lu G, Clark T, Siani-Rose MA, Ares M, Kulp DC, Haussler D: **Gene structure-based splice variant deconvolution using a microarray platform.** *Bioinformatics* 2003, **19 Suppl 1:**i315-i322.
24. Castle J, Garrett-Engele P, Armour CD, Duenwald SJ, Loerch PM, Meyer MR, Schadt EE, Stoughton R, Parrish ML, Shoemaker DD, Johnson JM: **Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing.** *Genome Biol* 2003, **4:**R66.
25. Neves G, Zucker J, Daly M, Chess A: **Stochastic yet biased expression of multiple *Dscam* splice variants by individual cells.** *Nat Genet* 2004, **36:**240-246.
26. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA: **Splicing graphs and EST assembly problem.** *Bioinformatics* 2002, **18 Suppl 1:**S181-S188.
27. Xing Y, Resch A, Lee C: **The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures.** *Genome Res* 2004, **14:**426-441.