Alternative splicing in eukaryotes: the norm, not an anomaly

Subhashini Srinivasan

As we enter the era of systems biology and personalized medicine, context-specific profiling of molecular genotypes and molecular phenotypes, such as single nucleotide polymorphisms, genes, proteins and metabolites, is becoming an integral part of medical research. However, molecular phenotype from alternative RNA splicing has been missing from this game because of the collective ignorance, lack of sensitive genomics technologies and inaccessibility of disease tissues critical for translation. This trend, however, is poised to change. It has now been shown using second-generation sequencing technologies that the majority of human genes express alternative spliced forms in a tissue, disease and individual-specific fashion[1], thus vastly expanding the repertoire of molecular phenotypes in humans.

In eukaryotic organisms, protein-coding regions are fragmented into several parts (exons) on the genome, which are separated by long stretches of non-coding regions (introns). The spliceosome handles the task of splicing out the introns and reassembling the exons together to form translatable mRNAs. It has been known for the last several decades that the spliceosome is capable of assembling exons alternatively to create distinct mRNA species. Considering that human genes, on an average, contain 7–8 exons, alternative splicing should have been anticipated to be the norm. But until the turn of the 21st century, alternative splicing was considered nature's anomaly, affecting less than 5% of the human genes. This assumption fostered the notion that the study of mRNA diversity is superfluous. The terms 'gene' and 'mRNA' have been used synonymously for nearly two decades. During the 1990s, when significant research was being done to predict the number of distinct genes coded by the human genome, the predictions varied widely depending on the scientific focus of the group. Incyte Genomics, a company specializing in RNA transcript sequencing, pegged the number of genes in the hundreds of thousands, which are several fold more than the estimations by groups working on DNA sequencing such as the Human Genome Project. Looking back, the RNA diversity from alternative splicing may explain much of the discrepancy in the estimation by the two groups.

The extent of RNA diversity from alternative splicing continued to be underestimated for lack of sensitive technologies[2,3]. The unexpectedly small number of genes revealed by the Human Genome Project forced us to look into RNA splicing to explain for the immense biological complexity in humans. As recently as 2007, a report from the ENCODE consortium, commissioned to study RNA complexity in the human transcriptome, revealed that the transcription process is far more complex than previously understood[4]. To our surprise, it has been shown that ~ 2% of the coding region of the human genome (exons) is actually scattered over 30% of the genome, thus increasing the intronic component of the genome an order of magnitude larger than the exonic regions. Since efficient recognition of cis-acting elements across introns by the spliceosome is critical to retain the integrity of the transcribed mRNA, splicing of long introns with tens of thousands of bases challenges the eukaryotic spliceosome and perhaps, is used to create RNA diversity from occasional slips. The spread of exons over a large swath of the genome also explains why large numbers of transcripts were found to unexpectedly bleed into neighbouring gene loci and why different genes overlapped on the same locus by transcribing distinct mRNAs from opposite strands. Advances in RNA sequencing (RNA-seq) technologies since 2007, have confirmed that alternative splicing of genes in human is the norm, not an anomaly[1].

Considering that most human diseases are tissue-specific and that majority of human genes express alternative spliced forms in a tissue-specific manner, the relevance of the combinatorial nature of function expansion in eukaryotes from alternative splicing can no longer be ignored. The most widely studied gene demonstrating the extent of function expansion resulting from alternative splicing mechanism is the *Drosophila* Down Syndrome Cell Adhesion Molecule (DSCAM). The DSCAM gene is capable of generating 38,016 distinct splice isoforms, which is twice the number of genes coded by the *Drosophila* genome. These isoforms include 19,008 distinct functional domains comprising unique combinations of multiple immunoglobulin-like folds. The preferential self-recognition of these domains is implicated in expanding the number of unique connections in the fly brain[5]. This may explain how the human brain may have evolved its complexity without a proportionate increase in the number of genes.

Reports on the relevance of alternative splicing in human diseases are accumulating in the literature. It has recently been shown that an N-terminal truncated splice variant of caboxypeptidase E induces tumour growth and is a predictive biomarker for metastasis[6]. Fifty per cent of disease-related mutations in DNA affect mRNA splicing in patients with neurofibromatosis type 1 (ref. 7). Just based on mutations at splice junctions alone, it is estimated that 15% of all disease-causing mutations may induce disrupting splicing. The actual percentage of disease-causing mutations that disrupt splicing ought to be much higher considering that mutations in cis-acting elements are not just localized to splice junctions. Apparently, many exonic mutations, including non-synonymous mutations, may alter splicing by disrupting exonic splicing enhancer (ESE) or suppressor (ESS) signatures[8]. Furthermore, the cis-acting elements are also spread across introns leading to intron splice enhancers (ISE) and silencers (ISS). However, the lack of bioinformatics tools to predict splicing changes resulting from mutations remains a major bottleneck in the assessment of the extent of disease-causing splicing defects resulting from point mutations[9].

Splicing defects caused by mutations also offer a novel therapeutic approach for treating genetic disorders. Muscular atrophy is linked to an ISS mutation in gene *SMN2*. An antisense oligonucleotide directed towards this mutation modifies the expression of the full-length *SMN2* gene in the nervous system, thus restoring the level of SMN2 expression

that could correct muscular atrophy[10]. A synthetic mRNA splicing modulator can diffuse into leaky muscle cells, modify splicing of DMD transcripts, induce the expression of partially functional dystrophin, and improve the function of some skeletal muscles[11]. Drugs or siRNAs that induce mitotic arrest promote proapoptotic splicing of Bcl-x, Mcl1 and caspase-9, and alter the splicing of other apoptotic transcripts[12]. Aberrant splicing in tumour cells correlates with tumerogenesis and targeting the RNA splicing machinery has been proposed as a novel treatment strategy[13].

Function expansion from alternative splicing is not only limited to humans, but is also common in plants. It has now been shown that 46% of rice genes are alternatively spliced[14]. This could be an underestimate, considering the diversity in rice cultivar and also that the estimate is limited to Japonica. The amylose content of rice from one cultivar to the other has been shown to depend on the expression levels of specific splice variants of Waxy gene[15]. Alternative splicing of members of the sulphate transporter family plays an important role in growth and development of rice and adaptation during stress conditions[16]. Mis-splicing in starch-synthesis genes in maize is also found to be crucial to controlling various grain traits[17].

The mounting literature establishing the relevance of alternative splicing to human disease and plant biotechnology demands a systematic, cost-effective approach to profiling RNA splicing in various biological contexts. Thanks to the advances in sequencing technologies in the recent years, sequencing of hundreds of millions of short RNA reads (RNA-seq) from biological samples is fast becoming affordable and routine. Data from RNA-seq experiments are inherently rich with information about both the molecular genotypes and phenotypes. The genome-wide profiling of RNA splicing is one of the pantheons of 'omics' information spewed by a single RNA-seq experiment. The redundancy in base coverage from RNA-seq experiments can also be harnessed to measure differential expression of RNA splice variants across contexts at resolutions yet not attainable by microarray technologies. Furthermore, considering the widely held belief that a majority of alternative splice events in humans and in various plants is yet to be discovered, the ability of RNA-seq to discover novel splice events is a huge plus.

Genome-wide profiling of RNA splicing in the biological context is moving from draught to surplus within a span of just few years. The affordability of RNA sequencing has also cleared the road for RNA-based signatures in the clinic. By sequencing genomes and transcripts from the same sample, context-specific RNA variants can now be associated with their respective causative genomic mutations using bioinformatics tools[18,19]. Efforts to translate RNA-based biomarkers into clinical diagnostics have led to the development of many sample-collection kits that would protect the RNA against enzymatic degradation at the point-of-care. Interestingly, the use of these kits has revealed that sufficient quantities of RNA can be extracted from cell-free saliva and other bodily fluids for transcriptome profiling[20]. At long last, mRNAs may make it to the ball (clinic) on time.

1. Wang, E. T. *et al.*, *Nature*, 2008, **456**, 470–476.
2. Modrek, B., Resch, A., Grasso, C. and Lee, C., *Nucleic Acids Res.*, 2001, **29**, 2850–2859.
3. Johnson, J. M. *et al.*, *Science*, 2003, **302**, 2141–2144.
4. Birney, E. *et al.*, *Nature*, 2007, **447**, 799–816.
5. Zipursky, S. L., Wojtowicz, W. M. and Hattori, D., *Trends Biochem. Sci*., 2006, **31**, 581–588.
6. Lee, T. K. *et al.*, *J. Clin. Invest.*, 2011, doi:10.1172/JCI40433
7. Buratti, E., Baralle, M. and Baralle, F. E., *Nucleic Acids Res*., 2006, **34**, 3494–3510.
8. Woolfe, A., Mullikin, J. C. and Elnitski, L., *Genome Biol*., 2010, **11**, R20.
9. ElSharawy, A. *et al*., *Hum. Mutat*., 2009, **30**, 625–632.
10. Burghes, A. H. M. and McGovern, V. L., *Genes Dev*., 2010, **24**, 1574–1579.
11. Aartsma-Rus, A., *RNA Biol*., 2010, **7**, 453–461.
12. Moore, M. J., Wang, Q., Kennedy, C. J. and Silver, P. A., *Cell*, 2010, **142**, 625–636.
13. Hayes, G. M., Carrigan, P. E., Beck, A. M. and Miller, L. J., *Cancer Res*., 2006, **66**, 3819–3827.
14. Lu, T. *et al*., *Genome Res*., 2010, **20**, 1238–1249.
15. Prathepha, P., *Pak. J. Biol. Sci*., 2007, **10**, 2500–2504.
16. Kumar, S., Asif, M. H., Chakrabarty, D., Tripathi, R. D. and Trivedi, P. K., *Funct Integr. Genomics*, 2011; doi:10.1007/s10142-010-0207-y
17. Ding, X. *et al*., *Plant Cell Rep*., 2009, **28**, 1487–1495.
18. Coulombe-Huntington, J., Lam, K. C. L., Dias, C. and Majewski, J., *PLoS Genet.*, 2009, **5**, e1000766.
19. Berger, M. F. *et al*., *Genome Res*., 2010, **20**, 413–427.
20. Zimmermann, B. G. and Wong, D. T., *Oral Oncol*., 2008, **44**, 425–429.

ACKNOWLEDGEMENTS. I thank the Department of Biotechnology, New Delhi and the Institute of Bioinformatics and Applied Biotechnology, Bangalore for the Ramalingaswamy Fellowship.

Subhashini Srinivasan is in the Institute of Bioinformatics and Applied Biotechnology, Biotech Park, Electronic City, Phase I, Bangalore 560 100, India.
e-mail: ssubha@ibab.ac.in