Alternative splicing and the evolution of phenotypic novelty

Citation for published version:

Digital Object Identifier (DOI):
10.1098/rstb.2015.0474

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Philosophical Transactions of the Royal Society B: Biological Sciences

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Alternative splicing and the evolution of phenotypic novelty

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Philosophical Transactions B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>RSTB-2015-0474.R1</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Review</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Bush, Stephen; University of Edinburgh, Roslin Institute Chen, Lu; Sichuan University West China Hospital, State Key Laboratory of Biotherapy Tovar Corona, Jaime; Wellcome Trust Sanger Institute Urrutia, Araxi; University of Bath,</td>
</tr>
<tr>
<td>Issue Code:</td>
<td>DIVERSITY</td>
</tr>
<tr>
<td>Subject:</td>
<td>Genomics &lt; BIOLOGY, Evolution &lt; BIOLOGY</td>
</tr>
<tr>
<td>Keywords:</td>
<td>alternative splicing, functional innovation, gene duplication, comparative genomics, effective population size, genetic drift</td>
</tr>
</tbody>
</table>
Alternative splicing and the evolution of phenotypic novelty

Stephen J. Bush\textsuperscript{1}, Lu Chen\textsuperscript{2}, Jaime M. Tovar-Corona\textsuperscript{3}, and Araxi O. Urrutia\textsuperscript{4,5*}

\textsuperscript{1}The Roslin Institute, University of Edinburgh, Edinburgh, EH25 9RG, UK.
\textsuperscript{2}West China Second University Hospital, State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610041, China.
\textsuperscript{3}Wellcome Trust Sanger Institute, Genome Campus, Hinxton, CB10 1SA, UK.
\textsuperscript{4}Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK.
\textsuperscript{5}Milner Centre for Evolution, University of Bath, Bath BA2 7AY, UK.

Keywords: Alternative splicing, functional innovation, gene duplication, comparative genomics, effective population size, genetic drift

Running Title: Alternative splicing & phenotype novelty

Summary

Alternative splicing, a mechanism of post-transcriptional RNA processing whereby a single gene can encode multiple distinct transcripts, has been proposed to underlie morphological innovations in multicellular organisms. Genes with developmental functions are enriched for alternative splicing events, suggestive of a contribution of alternative splicing to developmental plasticity. The role of alternative splicing as a source of transcript diversification has previously been compared to that of gene duplication, with the relationship between the two extensively explored. Alternative splicing is reduced following gene duplication with the retention of duplicate copies higher for genes which were alternatively spliced prior to duplication. Furthermore, and unlike the case for overall gene number, the proportion of alternatively spliced genes has also increased in line with the evolutionary diversification of cell types, suggesting alternative splicing may contribute to the complexity of developmental programmes. Together these observations suggest a prominent role for alternative splicing as a source of functional innovation. However, it is unknown whether the proliferation of alternative splicing events indeed reflects a functional expansion of the transcriptome or instead results from weaker selection acting on larger species, which tend to have a higher number of cell types and lower population sizes.

*Author for correspondence (A.Urrutia@bath.ac.uk).
Introduction

Since life first appeared on Earth, organisms have constantly changed and diversified to populate most environments. Although far outnumbered by single-celled organisms, multicellular organisms have evolved some of the most complex structures known, the consequence of elaborate and highly regulated programmes which allow their development from a single fertilised egg. The successive refinement of these programmes, in individual lineages over many generations, underlies the observed phenotypic diversity of the extant multicellular species. All observed phenotypic adaptations are, ultimately, encoded in the genome – but identifying which changes in the DNA underlie the evolution of novel phenotypes has proven to be a difficult task. Here we review the potential relevance of alternative splicing, a mechanism of post-transcriptional RNA processing, in the regulation of various biological processes and the evolution of novel phenotypes.

Alternative splicing is a process whereby multiple functionally distinct transcripts are encoded from a single gene by the selective removal or retention of exons and/or introns from the maturing RNA [1-4]. The process is highly regulated, involving trans-acting splicing factors and cis-acting regulatory motifs (see [5] for review) and so is susceptible to hereditary and somatic mutations [6]. Alternative splicing is common in many eukaryote lineages, including metazoans [7], fungi [8] and plants [9], with deep transcriptome sequencing of the human genome showing over 95% of multi-exon genes produce at least one alternatively spliced isoform [10, 11].

Alternative splicing has been found to be involved with numerous processes in multiple taxa including the virulence of pathogenic fungi [8], the flowering time of Arabidopsis [12], penile erection [13] and thermogenesis [14] in mice, and the neuronal development of primates [15, 16], rats [17], worms [18] and flies [19]. Many single-gene studies have also characterised the role of alternative splicing in various cellular processes (see [5, 20] for reviews), with a recent single-cell study detecting differential alternative isoform expression throughout the cell cycle [21]. Disruption or dysregulation of alternative splicing has also been associated with pathological states (see [22, 23] for reviews) further supporting a functional contribution of this process. aberrant alternative splicing has been linked to uncontrolled cell proliferation and chemotherapy resistance in cancer cell lines [24-28]. More broadly, genome-wide association studies in humans (which detect phenotype-predictive polymorphisms by genotyping affected and non-affected individuals) have uncovered a large number of intronic polymorphisms as significant predictors of disease susceptibility [29], many of which have been found to disrupt correct alternative splicing (e.g. [30-32]).

Some studies have associated alternative splicing with developmental plasticity and early speciation. Differential methylation patterns (which affect gene regulation) are associated with alternative splicing in bees which produce alternative castes [33]. Similarly, human body lice are thought to have emerged as an alternative morph of human head lice since the use of clothing became widespread. While there were no significant differences detected in which genes were expressed between the two morphs [34], morph-specific alternative splicing events were detected for over a thousand genes [35]. In particular, body lice specific alternative splicing events were enriched in genes associated with developmental gene ontology terms [35]. A comparable study of two mice subspecies that diverged about half a million years ago were also found to differ in their alternative splicing profiles [36].
Alternative splicing, functional innovation and the evolution of complexity in developmental programmes

Alternative splicing has been proposed to contribute to the evolution of novel phenotypes [37-40] as a small number of mutations (those in the 5’ or 3’ splice sites, for instance), can give rise to alternative splicing isoforms consisting of novel combinations of exons from existing genes [41, 42], providing an opportunity for the evolution of new functions. This has prompted comparisons between alternative splicing and gene duplication, widely recognised as a major source of phenotypic innovation [43, 44] (Holland, in this issue). The number of alternatively spliced isoforms per gene can decrease after a gene duplication event [45] suggesting that alternative splicing and gene duplication are to some extent equivalent mechanisms regulating transcript diversity. Several studies have reported that larger gene families (i.e. with more paralogues) have fewer alternative splicing events per gene compared to smaller families [46-48], further consistent with a reciprocal link. However, these correlations may also be explained by other factors, such as gene age (genes progressively acquire new splice variants over time, although the rate of increase is slowed by stronger purifying selection acting upon older genes) [49]. Figure 1 shows a model of the interplay between alternative splicing and gene
duplication.

Alternative splicing may also promote the retention of duplicated genes by facilitating subfunctionalisation [50] where gene copies acquire reciprocal mutations so that the duplicate copies fulfil the original’s gene function preventing them from becoming pseudogenes – the fate of most duplicate genes [51]. Single gene studies provide examples where duplicated copies are retained and diverge to resemble alternatively spliced isoforms produced by the ancestral single copy gene. For instance, in the vertebrate genome, which underwent at least one round of whole genome duplication early in its evolution [52, 53], there are three copies of the neuromuscular gene troponin I, each expressed in a distinct muscle type. These three genes are broadly equivalent to the alternatively spliced isoforms – also expressed in distinct muscle types – found in the urochordate Ciona intestinalis: a sister taxon to vertebrates, not affected by the whole genome duplication events, wherein troponin I is a single-copy gene [54]. Other examples of duplicate genes retained and evolving to resemble ancestral alternative splicing isoforms preceding gene duplication are characterised in the MITF [55], FoxP [56], and Syn-Timp families [57]. Moreover, two genome wide studies have shown that alternatively spliced genes are more likely to be found as duplicate copies in vertebrates [58] and fungi [59].

Soon after the publication of the human genome sequence – which revealed a lower than expected number of protein
coding genes – alternative splicing was proposed as a candidate to explain the diversification in the number of cell types observed in some eukaryotic lineages (a higher number of cell types in a given species is assumed to reflect increased organism complexity). As gene duplication has long been associated with functional innovation, it was initially assumed that overall gene number should correlate with the number of cell types [38, 60]. Gene duplication rates, however, failed to reflect the diversification of cell types observed in several eukaryotic lineages. There is a significant correlation between the two but weaker than expected and only moderately better if the analysis is restricted to metazoans [61].

Whole genome duplication events at the base of the vertebrate lineage [52, 53], which have the highest number of cell types among eukaryotes, have comparable numbers of genes to most invertebrates (that have not undergone whole genome duplication). The poor relationship between a species’ cell type diversity and its total gene number has become known as ‘the G-value paradox’ [60].
Although other genomic features have been shown to correlate with cell type number and may be important contributors to the evolution of complexity (i.e. [61, 62]), alternative splicing – as a mechanism allowing transcript diversification in the absence of increases in gene number – is a prime candidate to explain the G-value paradox [37, 38, 40, 63]. Comparative studies have reported marked differences in the prevalence of alternative splicing across eukaryotic lineages as well as a significant correlation between alternative splicing and the number of cell types per species [39, 61, 62, 64]. These results are in principle consistent with an adaptive role of alternative splicing in determining a genome’s functional information capacity and facilitating transcript diversification in species with greater numbers of cell types. It should be noted that some of the studies exploring changes in alternative splicing across evolutionary time and its association with cell type diversification have not corrected for unequal transcript coverage between species (which distort estimates of alternative splicing [37-39, 61, 64]) leading to inconclusive results [61, 64].

Functional molecular consequences of alternative splicing

It is clear that alternative splicing makes a large contribution to the number of distinct transcripts produced in many species – but the contribution of alternative splicing to the proteome is less clear [65]. Alternative splicing can modify the properties of encoded proteins, including the set of domains it contains, its binding properties, stability, intracellular localization and enzymatic activity [66-68]. Genes with higher levels of alternative splicing tend, in general, to have a higher number of protein-protein interactions [69]; those with alternative splicing events which are tissue specific also tend to have a more central position in PPI networks [70]. Furthermore, alternatively spliced regions preferentially encode residues found at the protein surface [71], which often contain interaction sites for proteins and their binding partners [72]. In addition, a gene’s alternatively spliced regions are more likely to encode disordered regions of protein (i.e. which have multiple stable configurations) [73]. These regions may also experience stronger selection for functions related to binding – flexible sensors are more adaptable than rigid ones [74, 75] – supporting an association of alternative splicing with the evolution of regulatory networks [76] (see [77] for a review of the regulatory circuits through which alternative splicing may act to regulate expression). Indeed, a comparison of the protein-protein interaction profiles of hundreds of protein isoform pairs suggests that the majority share less than half of their interactions – the principal and alternative isoforms belong to separate functional modules, suggesting functional differences between the set of splicing isoforms produced by the same gene [65]. Taken together, these observations are consistent with a role of alternative splicing in the diversification of the proteome and the modulation of protein-protein interactions.

Despite the various examples of alternatively spliced proteins having distinct functional roles, it is likely that a large proportion, or indeed the majority, of alternative splicing isoforms do not result in functional proteins [78-82]. An analysis of eight large-scale human proteomics experiments – collectively interrogating over 100 tissues, cell lines and developmental states with tandem mass spectrometry – found only 0.4% of the detected peptides are derived from alternatively spliced transcripts [83]. It is possible that the primary functional role of an alternative splicing event – if there is one at all – need not be at the protein level. For instance, alternative transcripts with ‘poison cassettes’ – exons with early in-frame stop codons – will never produce protein but instead down-regulate expression by diverting a proportion of pre-mRNA into the nonsense-mediated decay (NMD) pathway, a mechanism known as RUST (regulated unproductive splicing and translation) [84-86].
Selection versus genetic drift and the proliferation of alternative splicing

The increasing prevalence of alternative splicing in eukaryotes with higher number of cell types may be explained by neutral processes. In general, species with higher numbers of cell types also tend to have lower effective population sizes (the number of breeding individuals contributing to the next generation), which is thought to reduce the strength of selection relative to genetic drift [87, 88] (but see [89-91]). This is because in smaller populations there is a greater likelihood of neutral or slightly deleterious mutations – including those affecting splicing regulation – reaching fixation.

Thus, it is expected that the number of ‘noisy’ transcripts (those arising by erroneous splice site recognition) should be higher in species with lower effective population sizes resulting from mutations which introduce new but non-functional alternative splicing events, or that reduce the effectiveness of accurate intron splicing, increasing stochastic noise in RNA processing [78] – physicochemical fluctuations in the cellular environment that vary spliceosome efficiency [92]. As such, a reasonable null expectation is that many of the observed alternative splicing events are selectively neutral and that their existence is as a transient state, prior to being fixed or lost due to drift.

In the absence of experimental verification of each splicing event identified, the general functional contribution of splicing events is hard to assess. Under a drift model, it would be expected that the relationship between alternative splicing and cell type number should be explained by variations in effective population size. Our previous work suggests a strong relationship between alternative splicing and cell type number irrespective of effective population size [62], however, this finding is limited by the small quantity of effective population estimates available ($n = 12$). Although compendia of effective population size data have developed substantially [93, 94], there is still minimal overlap with the set of species for which comparative alternative splicing levels have been estimated. Taking adult body weight [95], average age at sexual maturity [96], and generation time [97] as proxies of effective population size [89] as well as ‘effective information’ (a value derived from principles of information theory as ‘the minimal amount of genomic information needed to construct an organism’) [98], the correlation of alternative splicing and cell type number remains significant [62] (Table 1 and Supplementary Table 1).

More complex species with higher number of cell types will also have longer generation times. If every time a genome is replicated there is a chance of error, then ‘complex’ species (which tend to have longer generation times and smaller effective population sizes), should have fewer replications per unit time [99]. We can speculate that consequently, mutations affecting splicing regulation – such as a single nucleotide polymorphism (SNP) that creates a new splice site [100] – are less likely to arise, but that new splice sites are also less likely to be eliminated (as the strength of drift relative to selection is higher). We can also predict that in ‘complex’ species, to disrupt, or create de novo, a splice site is to hit a large mutational target: the binding sites for proteins involved in exon recognition (necessary for accurate intron removal) are relatively large (approx. 30bp [78]), the binding sites for RNA-binding proteins (RBPs) are, although short, numerous [101], and the introns (in which new splice sites may arise) are longer in organisms with lower effective population sizes [88].

Another way to assess the functional contribution of alternative splicing is to examine the conservation of alternatively spliced events across evolution. A base-by-base analysis of the genotype-phenotype relationship for an alternatively spliced human exon demonstrated, that nucleotides which affect splicing regulation are dispersed the entire length of the...
exon and so any mutations could conceivably alter the accuracy of splicing [102]. This predicts the rapid divergence of alternative splicing events between species, except if they are selectively constrained. Conserved patterns of differential exon usage have been observed for a proportion of genes in comparisons of humans and mice [103-105], among primates [15, 106] and between grapevine cultivars [92]. Conserved alternative splicing profiles have also been identified in a lineage-, species- and tissue-specific manner among vertebrates [107, 108]. These studies suggest that in some of the species with the highest number of cell types, a significant number of the alternative splicing events are functional. However, whether the proportion of conserved alternative splicing events per species co-varies with the overall genome-wide prevalence of alternative splicing or with species complexity is unknown. In any case, increasing noise in alternative splicing regulation may, in itself, provide the substrate for a selective process and have adaptive significance to phenotypic plasticity, as it creates a ‘landscape of opportunities’ [109, 110].

It is likely that the observed distributions of alternative splicing events have been shaped by both adaptive and neutral processes, but further investigation is required to determine the evolutionary force of greatest significance.

Concluding remarks

The molecular drivers of development of an adult organism from a fertilised egg have been a matter of intense study over many decades. The discovery of DNA as the molecule containing the heritable information paved the way to many discoveries into the shared molecular pathways governing the development of a range of phenotypic traits. Advances in the ‘post-genomic era’ (characterised by its rapid accumulation of genomic, transcriptomic and proteomic profiling data) have allowed a better understanding of what molecular characteristics best explain the phenotypic differences between organisms. Perhaps more importantly, comparative studies are also helping to piece together the role of different aspects of genome evolution in the context of functional and phenotypic innovations. Alternative splicing provides an ideal mechanism for genomic novelty as it can result from a small number of mutations. Furthermore, while changes in the coding regions of genes will affect all transcripts containing this region, alternative splicing events can be expressed in highly specific conditions such as specific cell types or developmental stages. Here, we have summarised the evidence for a prominent role for alternative splicing in functional genomic evolution. Unlike gene number [61], alternative splicing is, at present, the foremost genomic predictor of increases in developmental complexity (using the number of cell types as proxy) [62]. It remains an open question as to what extent the association of alternative splicing and the number of distinct cell types is functional, or potentially a by-product of the fact that more complex organisms have lower effective population sizes and so are subject more strongly to drift.

Additional information

Data Accessibility
Primary data is included as supplementary material.

Authors’ Contributions
All authors were involved in the drafting of the article and approved the final version to be published.

Competing Interests
http://mc.manuscriptcentral.com/issue-ptrs
AOU is co-guest editor of this issue and is based at the same institution as the co-editor of the issue who was also the handling editor for this manuscript.

Acknowledgements
The authors wish to thank Professor Cheryll Tickle and three anonymous reviewers for their comments and suggestions which have greatly improved the quality of this manuscript.

Funding
This work was supported by CONACyT scholarships to ACM and JMTC, a UK-China scholarship for excellence and University of Bath research studentship to LC and Royal Society Dorothy Hodgkin Fellowship DH071902, Royal Society Research Grant RG0870644 and Royal Society Grant for Fellows RG080272 to AUO.

Additional information
Supplementary table 1. Assessment of the correlation between alternative splicing and organism complexity correcting for estimates and covariates of effective population size, including generation time, average age of sexual maturity and adult body weight.

References


93. Palstra F.P., Ruzzante D.E. 2008 Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? Molecular ecology 17(15), 3428-3447.


**Tables**

Table 1. Spearman Rho correlation coefficients between the average number of alternative splicing events per gene (ASL) and the proportion of genes alternatively spliced (ASP) with two measures of complexity (effective information and cell type number) correcting by various estimates and proxies of effective population size (Ne).

<table>
<thead>
<tr>
<th></th>
<th>ASL vs. Ne</th>
<th>ASL vs. complexity correcting by Ne</th>
<th>ASP vs. Ne</th>
<th>ASP vs. complexity correcting by Ne</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effective information</td>
<td>Cell type number</td>
<td>Effective information</td>
<td>Cell type number</td>
</tr>
<tr>
<td>Ne.μ, calculated using pi/π</td>
<td>0.22</td>
<td>0.90</td>
<td>0.88</td>
<td>0.27</td>
</tr>
<tr>
<td>Ne.μ, calculated using S</td>
<td>0.28</td>
<td>0.88</td>
<td>0.88</td>
<td>0.34</td>
</tr>
<tr>
<td>Ne.μ, calculating using eta/η</td>
<td>0.26</td>
<td>0.90</td>
<td>0.88</td>
<td>0.32</td>
</tr>
<tr>
<td>Adult body weight</td>
<td>0.64</td>
<td>0.21</td>
<td>0.93</td>
<td>0.39</td>
</tr>
<tr>
<td>Average age at sexual maturity</td>
<td>0.93</td>
<td>0.56</td>
<td>0.97</td>
<td>0.88</td>
</tr>
<tr>
<td>Generation time</td>
<td>0.70</td>
<td>0.86</td>
<td>0.78</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Notes. Significant correlation coefficients significant at p < 0.05 are marked in bold. See supplementary table 1 for underlying data and sources.
Figure caption

Figure 1. Model of the role of alternative splicing as a source of functional genomic innovation and retention of duplicate genes by facilitating sub-functionalisation. Novel alternative splicing events allow single-copy gene A to produce two protein isoforms, each optimised for expression in a specific tissue. After a duplication event the alternatively spliced exon is maintained in gene A2 but lost in A1. Divergence in the promoter areas of the duplicated genes lead to changes in gene expression profiles over time so that each gene is now only expressed in one tissue, mirroring the expression profile of the two isoforms of the original ancestral single-copy gene.