Microexons: novel regulators of the transcriptome

Ashton Curry-Hyde, Bei Jun Chen, James D. Mills and Michael Janitz

School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia; Department of (Neuro)Pathology, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; Paul-Flechsig-Institute for Brain Research, University of Leipzig, Leipzig, Germany

ABSTRACT

Alternative splicing of RNA is a fundamental post-transcriptional regulatory process that leads to a vast diversity of proteins being translated from a relatively small number of genomic loci. Microexons, a set of very small protein-coding sequences of 1-17 amino acids, have only recently been recognised as an important part of pre-mRNA processing. Recent studies have revealed that microexons can play important roles in various cellular functions, protein-protein interactions and have also been associated with various neurological diseases. This review provides an update on research covering the functional impact of microexons on the biology of a cell and disease, and the mechanisms by which their splicing is regulated. Finally, the current bioinformatics methods for detecting microexons are discussed.

ARTICLE HISTORY
Received 11 August 2017
Accepted 13 February 2018

KEYWORDS
Microexon; genome; transcriptome; proteome; RNA-Seq; alternative splicing; VAST-TOOLS

Introduction

Every cell in the human body contains the exact same genetic information, but each cell varies in the parts of the information it uses to carry out its functions. It has long been established that the complexity of an organism is not directly proportional to the number of genes present. This means that the complexity observed in higher organisms, such as humans, must stem from other sources or processes. Alternative splicing is the process by which the exons of primary transcripts (pre-mRNAs) can be spliced into different arrangements to produce structurally and functionally distinct mRNA and protein variants [1]. It is through the process of alternative splicing (AS) that the variety of cellular functions, such as co-existence of enzymatic isoforms within tissues and organs, are achieved [2]. Exonic sequences are the regions that are intrinsic to the daily function of organisms through the central dogma of molecular biology: deoxyribonucleic acid (DNA) undergoes transcription to become ribonucleic acid (RNA), which undergoes translation to produce a functional product known as a protein [3]. It is these proteins that ultimately allow organisms to function in day-to-day life, and are determined through the mechanisms of AS which combines exonic sequences to derive specific proteins.

Exons that are smaller than 51 nucleotides (nt) long are called microexons [4,5]; they are more prone to exon skipping as it has been shown in wide range of vertebrates [6,7]. The maximal length of 51 nt for microexons has been set following the study by Li et al. reporting splicing of microexons sized from 3-51 nt long [8]. Microexons were initially identified in the 1980’s, by a number of different research labs [4,9–13]. However, due to their unknown biological significance and the technical difficulties associated with identification, microexons were largely considered as genetic noise and by-product of splicing [14]. In 2003, a novel and remarkably successful computational method was developed for microexon discovery in entire genomes [15].

This review provides an update on recent advances in research on microexons. We particularly focus on biological function of microexons inclusion and their role in molecular pathology of complex diseases. Recent views on the mechanisms by which microexon splicing is regulated are also covered.

Alternative splicing

Alternative splicing (AS) is a dynamic, well-organised mechanism that enhances the complexity of the
eukaryotic transcriptome and proteome by increasing the number of mRNA variants produced from an individual gene [5,16,17]. AS plays critical roles in a diverse range of cellular functions such as stem cell pluripotency maintenance and tissue-specific protein maintenance in the brain and muscle. Transcript isoforms are expressed through AS of multi-exon genes, expanding functional diversity [5,17]. AS misregulation can be catastrophic for an organism, playing roles in cancer development and also in a variety of neurological diseases [4].

AS occurs via the spliceosome, an RNA-protein complex. The spliceosome assembles at both 5' and 3' ends of intronic sequences after scanning for the constant splicing sites, GU and AG, respectively [15]. AS of precursor mRNA (pre-mRNA) by the spliceosome results in various selections of 5' of precursor mRNA (pre-mRNA) by the spliceosome ends of intronic sequences after scanning for the consensus sequence [16]) that are enriched by enhancer elements positively affect the efficiency of splicing, in addition to the size of introns and exons. In higher eukaryotes, specifically mammals, exons are ~10X shorter than introns, with 80% being smaller than 200 nt [19]. The difference in sizes makes intron definition a challenging prospect, and led to the exon-definition model proposition which recruits splicing machinery through coordinated recognition of splicing sites in exons, ensuring a range of optimally sized exons are efficiently spliced by the spliceosome. Despite the ideal modelling of exon definition, exons shorter than 51 nt still prove problematic to define.

In vertebrates, recognition of the initial splice site in pre-mRNA might potentially be inefficient when larger intronic regions encompass small exons [20]. Exon bridging or definition events facilitate smaller exon, paired splice site recognition. Early on in the assembly of the spliceosome, recognition of an exon occurs through interactions between U2 small nuclear ribonucleoprotein (snRNP) auxiliary factors bound to the pyrimidine tract of the 3' splice site and U1 snRNPs bound to the 5' splice site [20]. Intronic enhancers adjacent to exons are mandatory for efficient exon recognition and splicing. A core repeat unit within an ISE located adjacent to the 5' splice site of a 6 nt microexon from the chicken cardiac troponin T gene (cTNT) was reported along with identification of a protein recognising the core sequence. Within the 130 nt long ISE were six repeats of the G-rich sequence: GGGGCUG. The mammalian splicing factor 1 (SF1) binds to this repeated sequence, via UV-cross-linking experiments, thus contributing to the identification of microexons in the exon-definition model [20].

**Biogenesis and function of microexons**

Microexons lack the required exonic splicing enhancers, and due to their size the splicing machinery cannot assemble at both the 5' and 3' splice sites [5]. Two independent groups published groundbreaking results from RNA-seq data from several species, revealing highly conserved microexons that display features specific for regulating their inclusion, and the impact of these microexons on protein function in neurogenesis and other brain functions [8,21].

Irimia et al. developed a module that systematically analysed and defined all neural-regulated splicing patterns from >100 humans and mouse tissue types, focussing on microexons with a particularly short length (3-15 nt) [21]. Microexons comprised 1/3 of conserved neural-regulated splicing between human and mouse samples. The inclusion of identified neural microexons in the final transcript is regulated by nSR100 (also known as SRRM4 – serine/arginine repetitive matrix protein 4 [22]), a brain-specific RBP, which binds to the intronic enhancer UGC motifs located near the 3' splice sites, and also promotes the splicing of microexons in neurogenesis. These neural microexons have lengths that are multiples of 3 nt, which means the microexon is in-frame, and can likely produce an altered protein isoform. Experimental evidence showed that microexon regulation is activated during the later stages of neural differentiation and that the inclusion of microexons during AS events enhances protein-protein interactions [23].

A survey of over 900 human and mouse tissue samples, largely brain tissue, revealed that microexons are alternatively or constitutively spliced, they exhibit tissue-specific pattern of inclusion, and they are evolutionarily conserved. Additionally, it was determined
that microexons which demonstrate regulation through RBPs, such as RBFox, alter protein sequences, which ultimately alters the methods of protein-protein interactions [8]. Li et al. regarded microexons as sections between annotated splice junctions, and AS microexons interacted more with the splicing machinery than CS microexons [8]. Tissue-specific RBFox proteins play critical roles in brain-specific microexon enhancement [8]. Both studies established a distinct inverse relationship between the tendency for an alternative exon to be included in neural cells and tissues, and its length sized between 3-15 nt long [8,18,21].

In AS, microexons have compensatory mechanisms that combat the primary disadvantages they suffer because of their size, allosteric hindrance based on the exon-definition model, and the difficulty ensued to include the required ESE's in the spliced sequence. The misregulation of AS is becoming increasingly implicated in neurodegenerative disorders such as amyotrophic lateral sclerosis (ALS) [22].

Analysis of regulatory features in 7949 microexons in the brain revealed inefficient exon definition could be compensated for by more than just efficient intron definition; 3’ and 5’ splice sites are significantly stronger in CS microexons compared to AS microexons and longer exons; ESE density in CS microexons is greater compared to longer CS exons, potentially due to their shorter sequence length [8]. Proximally to the microexon splice sites, it was also observed that certain motifs co-occurred in addition to highly enriched uridine and cystidine elements, located 10-20 nt upstream of the 3’ splice site compared to constitutive exons [8]. Some evidence exists highlighting some sequence motifs that promote the inclusion of microexons in vertebrates.

Holter et al. identified two novel GABA receptor variants in the metabotropic GABABR1 gene that resulted from AS events in the gene, specifically centred on exon 4 as the locus for regulated splicing in GABABR1 [24]. GABA is an inhibitory neurotransmitter in mammalian brains and GABABR1 contains multiple splice sites in rats (8 splice sites) and humans [24]. Exon 4 flanking sequences contain strong splice sites and multiple ESEs in the 80bps upstream of the 5’ splice site, but also lack the 5’ terminal G nucleotide, which could potentially weaken the 3’ splicing site, which further adds to experimental evidence involving alternative exon determination being related to enhancer elements and splice site strength [24].

During development and cellular differentiation, microexons undergo dynamic changes, particularly in the central nervous system (CNS). AS in neuronal microexons is particularly high compared to other cell types within the human CNS, and more than 90% of regulated microexons have their highest inclusion levels in neurons [21]. Comparisons of AS events in human and mouse suggest that 1/3 of neuron-specific AS is represented by microexons. It is believed the AS factors RBFox, PTBP1 and nSR100 contribute to microexon tissue/cell-type specific inclusion regulation [8,21]. The extent of the effect of microexon alterations in neurological disorders is still unclear [18,24].

In their recent review Ustianenko et al. highlight that microexon function is determined by the position of the spliced exon along the mRNA sequence. If a
microexon is in-frame and does not contain a stop codon, the amino acid sequence is entirely altered [4]. Hence the properties of the translated protein can change dramatically. For example, inclusion of microexons will result in altered protein binding properties or alternative pattern of post-translational modifications (Figure 1). If the exon is not in-frame, it will generally result in a premature stop codon, resulting in early degradation of the transcript by nonsense-mediated decay (NMD) [4]. 80-90% of microexons maintain an open reading frame as their length is a multiple of 3 [8,21]. The microexons that do not fit into this category will likely cause a frame-shift in the RNA sequence, causing the mRNA surveillance pathway to trigger NMD, in addition to triggering regulatory mechanisms designed to maintain a steady-state gene expression level.

In 2014 the largest database of neural-specific AS events was developed. Microexons 3-15 nt in length were reported to convey the most striking evolutionary conserved and switch-like regulation, in addition to regulating the interaction of protein domains involved in the process of neurogenesis [21]. Neuronal microexon encoding amino acid sequences show altered protein domain architecture and structure, modifying protein-protein interactions in specific tissues. Microexons are typically the central nodes in protein interactions and protein complexes [4,21]. An example of altered protein function protrudes in (ZFYVE27), where a 21 nt microexon is included in neurons, but not oligodendrocytes, which encourages neurite outgrowth and axon growth, alongside neuronal polarity. Microexons also create sites for the post-translational modification (PTM) of proteins [4].

Microexons and disease

Misregulation of AS is frequently caused by mutations occurring at the 5’ or 3’ splice sites, the branch point, or the spliceosome recruitment regulatory elements including ESE, ISE, ESS and ISS [19]; this leads to perturbation of key splicing regulators, or disruption of RBP binding to mutated splice sites. Splicing misregulation is recognised as the cause of a number of diseases such as autism spectrum disorder (ASD), spinal muscular atrophy [13], myotonic dystrophy, cancer, frontotemporal dementia and ALS [4,25].

Xiong et al. focussed on the effect of missense single nucleotide variations (SNVs) on splicing regulation as opposed to the effect on protein function [25]. Investigation of RNA splicing misregulation revealed missense SNVs related to both the intronic and exonic regions that are causative of disease. Missense SNVs related to intronic disease were nine times more likely to disrupt splicing regulation compared to SNPs if they are within 30 nt from the splice site. In regard to exonic disease, missense SNVs are 9.3 times more likely than SNPs to disrupt splicing regulation. Xiong et al. chose to investigate SNVs rather than SNPs, as <5% of SNVs determined through genome wide association studies (GWAS) are estimated to cause the same misregulation effect as SNVs. This groups methodology maps SNVs to exonic and intronic sequences that contain regulatory code for exons in genes; 17.45% of SNVs studied were largely rare SNPs with a minor allele frequency (MAF) of <1%. The effect of each specific splicing regulation in a sequence was scored according to a regulatory model with and without the presence of the SNV, providing a difference in predicted splicing level for each tissue [25].

ASD has been linked to neural-specific AS misregulation in microexons [21], specifically RBFOX1-dependent AS [5]. Almost 30% of AS microexons in the brains of individuals with ASD were misregulated, a percentage that is directly correlated to the level of nSR100 expression levels in the same patients [21,23]. Extensive AS pattern deviation in cortical regions of the brain in ASD is related to splicing misregulation in ASD genes (e.g. neurexin and neuroligins), in addition to evidence shown by transcriptomic profiling in cases of ASD [25]. The study found an association between the reduction in the level of the neuronal-specific splicing factor nSR100, which regulates neural microexons by binding to ISE motifs, and microexon misregulation in the brains of people with ASD.

The large glycoprotein, reelin, is critical in neural development. Reelin is secreted in the extracellular matrix of cells in the brain and is >388 kDa in size [26]. Two evolutionary conserved AS events in the 3’ end of the mouse reelin mRNA in the mouse model were observed. First, a 6-nucleotide, brain-specific microexon is skipped in about 10% of reelin RNA. In addition, an alternative polyadenylation event involving 10-25% of reelin mRNA results in secretion of a truncated protein lacking the terminal, highly basic stretch. Interestingly, the same reelin mRNA modifications were also observed in human, turtle, lizard and rat brain tissue. Reelin is believed to act on target neurons via the extracellular matrix, requiring the response of the Dab1 gene product, however, the expression of reelin in other structure and also in its alternative forms are suggestive of it having various functions [26].
Microexons are shown to play key roles in protein-protein interactions and PTM, and therefore are potential candidates affected by the perturbations derived from AS misregulation [4]. The majority of microexons play a distinctive role in altering protein-protein interactions, yet some also contribute novel platforms or newly charged regions that allow PTM. It is highly speculated that microexons also influence other tissues, which would further contribute to the profile of disease that microexons influence [5]. Proteomic technologies have proved a wealth of information regarding protein relations and function, including logging PTMs across species. Specific protein modification sites and the specific regulatory roles critical to the function of the protein can be better documented by classifying PTMs according to their degree of conservation, creating a more sound picture of biochemical and biological evolutionary relationships [27].

Identification of microexons

The revolutionary procedure determined by the Salzberg lab in 2003 [15] uses a reference genome and unaligned segments of the sequence of interest with recognised flanking splice signals to search for a match for the unaligned segment in the reference genome. When this algorithm provides a perfect match for the unaligned segment, a novel microexon has been found [4]. Enhancement of this method by Wu and Watanabe involved the addition of statistical significance scores for the microexons and a new detection algorithm with GMAP (EST alignment tool) [28].

The next major progressive step in microexon discovery occurred in 2014 when Irimia et al. published a vertebrate alternative splicing and transcription tools (VAST-TOOLS) software program was developed [21]. VAST-TOOLS builds an exon-microexon-exon junction database from recognised intronic regions in cDNA libraries, from which possible microexons are computed in silico by searching through splice sites that are 3-15 nt apart. RNA-seq library reads are aligned to give evidential support of novel microexon identification in the exon-microexon-exon junction database [4,21]. Using VAST-TOOLS 696 AS microexons of 3-27 nt in length range was identified [21].

The third bioinformatics tool used to identify microexons is called ATMap [8]. ATMap aligns RNA-Seq reads to a transcript model library to search for small insertions 3-51 nt long that are possibly novel microexons, using recognised splice sites surrounding the unmapped region in question. Recognised high confidence junctions based on splice site motifs are defined by comparing these small insertions with a reference genome, generating inclusion and exclusion isoforms to quantify microexons [4]. If the reads map correctly according to the reference genomes, microexon/s are identified [4,8]. Utilization of ATMap led to identification of 13,145 constitutive and AS microexons 6-51 nt using ATMap. The same study also found that the ATMap approach gives similar results to the RNA-seq alignment software, OLEgo [8].

Conclusions and perspectives

The discovery of microexons, and the subsequent methodology to identify and analyse these sequences as functional components of the human genome, is pinnacle in the research focussed on greater understanding of cellular function, protein-protein interactions, diseased states, but also in discovering an additional element to the human genome which had previously been overlooked. The methods recently developed in several laboratories have been crucial in our present understanding of microexons biological significance. While the advancements in microexon research is formidable, further studies are required to incorporate a diverse range of cell type and diseased states, including neurological disorders, with a focus on microexon contributions to protein alteration.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Michael Janitz  http://orcid.org/0000-0002-4878-0259

References

[1] Regulation Cfg. Center for Genomic Regulation. Microexons: Small fragments of genes, essential for neuron maturation. 2014 [cited 2014 Dec 18]. ScienceDaily. Available from: http://www.sciencedaily.com/releases/2014/12/141218131525.htm
[2] Dvinge H, Kim E, Abdel-Wahab O, et al. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer. 2016;16:413–430.
[3] Crick F. Central dogma of molecular biology. Nature. 1970;227:561–563.
Ustianenko D, Weyn-Vanhentenryck SM, Zhang C. Microexons: discovery, regulation, and function. Wiley Interdiscip Rev RNA 2017.

Yang L, Chen LL. Microexons go big. Cell. 2014;159:1488–1489.

Carlo T, Sterner D, Berget S. An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. CARLO 1996.pdf. RNA. 1996;2:342–353.

Dominski Z, Kole R. Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol. 1991;11:6075–6083.

Li YI, Sanchez-Pulido L, Haerty W, et al. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res. 2015;25:1–13.

Beachy PA, Halfand SL, Hogness DS. Segmental distribution of bithorax complex proteins during Drosophila development. Nature. 1985;313:545–551.

Cooper TA, Ordahl CP. A single cardiac troponin T gene generates embryonic and adult isoforms via developmentally regulated alternate splicing. J Biol Chem. 1985;260:11140–11148.

McAllister L, Rehm EJ, Goodman GS, et al. Alternative splicing of micro-exons creates multiple forms of the insect cell adhesion molecule fasciclin I. J Neurosci. 1992;12:895–905.

Santoni M, Marthels D, Vopper G, et al. Differential exon usage involving an unusual splicing mechanism generates at least eight types of NCAM cDNA in mouse brain. The EMBO Journal. 1989;8:385–392.

Small SJ, Haines SL, Akeson RA. Polypeptide variation in an N-CAM extracellular immunoglobulin-like fold is developmentally regulated through alternative splicing. Neuron. 1988;1:1007–1017.

Scheckel C, Darnell RB. Microexons—tiny but mighty. EMBO J. 2015;34:273–274.

Vollovsky N, Haas BJ, Salzberg SL. Computational discovery of internal micro-exons. Genome Res. 2003;13:1216–1221.

Kornblitt AR, Schor IE, Allo M, et al. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol. 2013;14:153–165.

Gueroussov S, Gonatopoulos-Pournatzis T, Irimia M, et al. An alternative splicing event amplifies evolutionary differences between vertebrates. Science. 2015;349:868–873.

Raj B, Blencowe BJ. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron. 2015;87:14–27.

Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17:19–32.

Carlo T, Sierra R, Berget SM. A 5′ Splice Site-Proximal Enhancer Binds SF1 and Activates Exon Bridging of a Microexon. Molecular and Cellular Biology. 2000;20:3988–3995.

Irimia M, Weatheritt RJ, Ellis JD, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014;159:1511–1523.

Vuong CK, Black DL, Zheng S. The neurogenetics of alternative splicing. Nat Rev Neurosci. 2016;17:265–281.

LaFlamme B. Microexons on the brain. Nature Genetics. 2015;47:105.

Holter J, Davies J, Leresche N, et al. Identification of Two Further Splice Variants of GABAB1 Characterizes the Conserved Micro-Exon 4 as a Hot Spot for Regulated Splicing in the Rat Brain. Journal of Molecular Neuroscience. 2005;25:99–108.

Xiong HY, Alipanahi B, Lee IJ, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347:1254806.

de Rouvroit CL, Bernier B, Royaux I, et al. Evolutionarily Conserved, Alternative Splicing of Reelin during Brain Development. Experimental Neurology. 1999;156:229–238.

Greco TM, Cristea IM. The Biochemical Evolution of Protein Complexes. Trends Biochem Sci. 2016;41:4–6.

Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–1875.