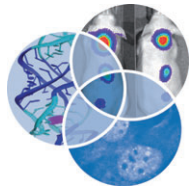
Advanced Review



On the path to genetic novelties: insights from programmed DNA elimination and RNA splicing

Francesco Catania[1]* and Jürgen Schmitz[2]

Understanding how genetic novelties arise is a central goal of evolutionary biology. To this end, programmed DNA elimination and RNA splicing deserve special consideration. While programmed DNA elimination reshapes genomes by eliminating chromatin during organismal development, RNA splicing rearranges genetic messages by removing intronic regions during transcription. Small RNAs help to mediate this class of sequence reorganization, which is not error-free. It is this imperfection that makes programmed DNA elimination and RNA splicing excellent candidates for generating evolutionary novelties. Leveraging a number of these two processes' mechanistic and evolutionary properties, which have been uncovered over the past years, we present recently proposed models and empirical evidence for how splicing can shape the structure of protein-coding genes in eukaryotes. We also chronicle a number of intriguing similarities between the processes of programmed DNA elimination and RNA splicing, and highlight the role that the variation in the population-genetic environment may play in shaping their target sequences.

How to cite this article:
WIREs RNA 2015, 6:547–561. doi: 10.1002/wrna.1293

INTRODUCTION

The elucidation of the processes that govern the evolution of genomes and gene architecture is a central challenge facing biologists. Although a variety of experimental and bioinformatics tools currently enable us to detect and characterize newly emerged genetic structures with ease, less facile is the task of understanding the mechanisms contributing to the formation of these genetic novelties. In this article, we focus on programmed DNA elimination and RNA splicing. Verifiable hypotheses concerning universal mechanisms for the emergence of novel gene structures surface from our improved understanding of these two mechanistically, phylogenetically, and functionally distinct processes. In particular, there is evidence to suggest that, in contrast to novelty-generating mechanisms such as transposon insertions or DNA replication errors, programmed DNA elimination and RNA splicing may lead to the emergence of evolutionary innovations via an analogous process of mutual conversion between sequence units that are commonly considered to be evolutionarily and functionally separate (i.e., germline vs somatic sequences; exonic vs intronic sequences).

DNA-LEVEL SPLICING

Genomes are highly dynamic structures. Not only do they undergo changes in content and architecture over evolutionary time, they can also experience less or more dramatic rearrangements within a lifetime.[1] Regardless of the temporal scale, DNA-level splicing is a major contributor to genome repatterning. This molecular process consists of two steps: (1) the recognition and excision of internal regions of DNA, and (2) the rejoining of the flanking sequences. Such

*Correspondence to: francesco.catania@uni-muenster.de

[1]Institute for Evolution and Biodiversity, University of Münster, Münster, Germany

[2]Institute of Experimental Pathology (ZMBE), University of Münster, Münster, Germany

Conflict of interest: The authors have declared no conflicts of interest for this article.

cut-and-join reactions are commonly associated with the movement of DNA transposons, which may relocate from one position on the genome to another at any time during an organism's life.[2] In addition, cut-and-join reactions can be associated with events of gene or genome reorganization, which do not involve transposition and take place at specific times during development in several eukaryotic lineages.[3,4] One typical example of such DNA-level splicing occurs in jawed vertebrates during the differentiation of embryonic stem cells into lymphocytes. During V(D)J recombination, various reorganizational events including a process of DNA-level splicing known as programmed DNA elimination shape the immunoglobulin genes and T-cell receptor genes and are crucial for the formation of antibodies.[5]

Programmed DNA Elimination: How It Works

Programmed DNA elimination is both extensive and well studied in ciliated protozoa, two aspects that make these single-celled organisms excellent models for gaining further insights into this process. Ciliates are characterized by the presence of two functionally distinct types of nuclei: a diploid germline micronucleus (MIC) that produces gametic nuclei during sexual events, and a polyploid somatic macronucleus (MAC) that is expressed throughout the organism's vegetative life. Programmed DNA elimination in ciliates takes place during sexual reproduction when the new MAC is regenerated from a mitotic copy of the zygotic nucleus—while the maternal MAC is lost. This DNA-level splicing process is critical for the development of a functional MAC genome in that it guarantees the precise elimination of germline-specific DNA sequences that frequently interrupt coding sequences. These spliced sequences are known as internal eliminated sequences (IESs).[6] IES excision in ciliates is particularly well studied in Paramecium,[7–9] Tetrahymena,[10–12] and Oxytricha.[13–15] Here, we will focus on the process of IES excision in the best-studied species of Paramecium, *Paramecium tetraurelia*.

The ~45,000 known IESs in *P. tetraurelia* are typically unique, short (93% are shorter than 150 bp), AT-rich (~80%), and frequently reside in coding regions (~77%).[16] Additionally, *P. tetraurelia* IESs are invariably flanked by 5′-TA-3′ dinucleotides, one of which is retained in the somatic genome subsequent to IES excision. These TA dinucleotides are part of larger (8-bp) imperfect inverted terminal repeats, whose consensus (5′-**TA**YAGYNR-3′) is similar to that of Tc1-related DNA transposons (5′-**TA**CAGTKS-3′).[17] This latter resemblance taken together with the observation that PiggyMac, a domesticated PiggyBac-related transposase, is required for IES excision in *P. tetraurelia*,[18] and the identification of some IESs which clearly originate from transposons,[16] has led to the hypothesis that IESs in Paramecium have evolved from ancestral germline insertions of DNA transposons.[19] Nevertheless, other potential sources for the origin of IESs cannot be entirely ruled out (see below).

Both genetic and epigenetic mechanisms control IES excision in *P. tetraurelia*. The disruption of the 5′-TA-3′ dinucleotide or mutations at sites within the terminal repeat result in the retention of IESs in the MAC genome, suggesting that these sites serve as recognition or excision signals.[20–22] Additionally, the microinjection of IESs into the maternal (old) MAC causes the retention of a subset (~1/3) of homologous IESs in the developing MAC.[23] Interestingly, retained IESs may inhibit the elimination of homologous IESs from the MAC in successive sexual generations, as long as their terminal repeats, or a part of their sites at least, remain unchanged.[22] Collectively, these observations imply that while dependent on the IES terminal sequences, the elimination of a fraction of IESs from the macronucleus of sexual progeny in Paramecium is mediated by epigenetic, homology-dependent, mechanisms.

Numerous experimental observations have led to the formulation of a model—the *scnRNA* or *genome-scanning* model—for the excision of epigenetically (or maternally) controlled IESs.[24,25] Small noncoding RNA molecules play a key role in this model, which consists of three main phases. First, 25-nt small RNAs (known as scnRNAs) are produced in the MIC following genome-wide bidirectional transcription and Dicer-like-protein-mediated cleavage of the resulting double-stranded RNA transcripts. Second, Piwi-bound scnRNAs are transported to the maternal MAC where a subtraction process takes place. In essence, the scnRNAs that are able to pair with maternal long noncoding RNAs are inactivated, whereas the germline-specific (IES-matching) scnRNAs remain unpaired and intact. Third, these unpaired scnRNAs are transported to the developing MAC, where they facilitate IES excision by PiggyMac via mechanisms involving chromatin modifications.[8] An addendum to this model has been recently proposed after a distinct class of small RNAs—termed iesRNAs—has been found to participate in the development of the Paramecium macronuclear genome.[9] iesRNAs affect the excision of a fraction of nonmaternally controlled IESs.

The Imperfection of Programmed DNA Elimination and Its Evolutionary Significance

The availability of the MAC genome first[26] and of (large part of) the MIC genome later has enabled the first large-scale detection and characterization of IESs in Paramecium.[16] At present it is clear that while largely faithful, as any other biological process programmed DNA elimination is imperfect, which makes it evolutionarily significant.

In addition to DNA-level splicing, the development of a new macronuclear genome in Paramecium entails chromosome fragmentation and genome amplification.[27,28] Because genome amplification precedes IES removal during the sexual event, identical MAC chromosomes may be processed differently in the same cell. In line with this, a study of the macronuclear DNA sequence variability in a homozygous strain of *P. tetraurelia* revealed hundreds of loci that are partially mapped by low-frequency reads with a TA-flanked insertion.[29] The vast majority of these insertions were later shown to be (incompletely excised) IESs.[16] Moreover, partially retained TA-flanked sequences (i.e., putative IESs) were also found in the macronucleus of *Paramecium biaurelia* and *Paramecium sexaurelia*, two species that are closely related to *P. tetraurelia*.[30] These studies strongly suggest that IESs in Paramecium may undergo incomplete excision during the formation of the new MAC genome.

Imperfect programmed DNA elimination involves not only IESs (i.e., germline-specific sequences) but also soma-specific sequences. The very studies that detected putative imperfect IES excisions also uncovered hundreds of loci in the Paramecium macronuclear genome, which are mapped by low-frequency reads with a TA-flanked deletion, rather than a TA-flanked insertion. These excised macronuclear DNA regions contain terminal sequences that resemble the terminal repeats of true IESs and, hence, it is plausible that their excision also results from the erroneous recognition by PiggyMac. In sum, both IESs and macronuclear DNA regions can be imperfectly excised at each sexual event in Paramecium.

Besides being imperfectly excised, IESs (and macronuclear DNA regions) might occasionally be completely retained in (or completely excised from) the newly developing somatic genome of Paramecium. In particular, the results of a recent comparative genomic analysis of *P. tetraurelia*, *P. biaurelia*, and *P. sexaurelia* revealed that putative IESs or TA-flanked macronuclear DNA sequences present in (or absent from) the macronuclear genome assembly of one species were absent from (or retained in) the MAC genome of one of the two remaining species under study.[30] The detected cases of differential IES retention/macronuclear DNA excision suggest that germline-specific sequences contribute to and diversify the somatic genome contents in these species. Additionally, retained IESs and spliced macronuclear DNA regions not infrequently reside in coding sequences, thus potentially having effects on fitness.

Programmed DNA Elimination and the Birth of Genetic Novelties

The observations described above offer a static view of the molecular condition of the Paramecium MAC genome and suggest that like erroneous (incomplete) elimination of macronuclear DNA regions, unfaithful IES excision also generates distinct DNA variants. These observations give rise to the question: Does imperfect DNA-level splicing have evolutionary consequences? Recent studies have shown that erroneous programmed DNA elimination has a selective cost (see below). This implies that rearrangement errors at a given locus may occasionally encompass a fraction of macronuclear copies that are sufficiently large to have negative effects on fitness and to, thus, trigger a selection response. Intriguingly, the increase in frequency that may facilitate the purging of deleterious DNA splicing variants might also favor the permanent integration of splicing variants in the MAC genome. It has been shown that sufficiently large fractions of DNA variants in the macronucleus facilitate the transmission of homologous variants to the next sexual generation through epigenetic mechanisms.[31] Thus, it is entirely possible that nonlethal and heritable somatic DNA variants may ultimately spread through a sexual population with some probability of fixation (see Figure 1). One fulfilled prediction of this hypothetical evolutionary scenario is that the IES excision profile differs to some extent between Paramecium species.[30]

Within a cell, changes in isoform frequencies over subsequent sexual generations may result from the optimization or the weakening of *cis*-acting signals that modulate programmed DNA elimination. Although it is currently unclear how the quality of these signals (or what precisely these signals are for that matter) may be assessed, it is clear that some *cis*-acting sequences affect IES recognition and excision.[32] It is reasonable then that the progressive accumulation of germline mutations that weaken *cis*-acting DNA splicing signals facilitates an increased frequency of IESs in the macronucleus over evolutionary time. Similar to splicing-disruptive

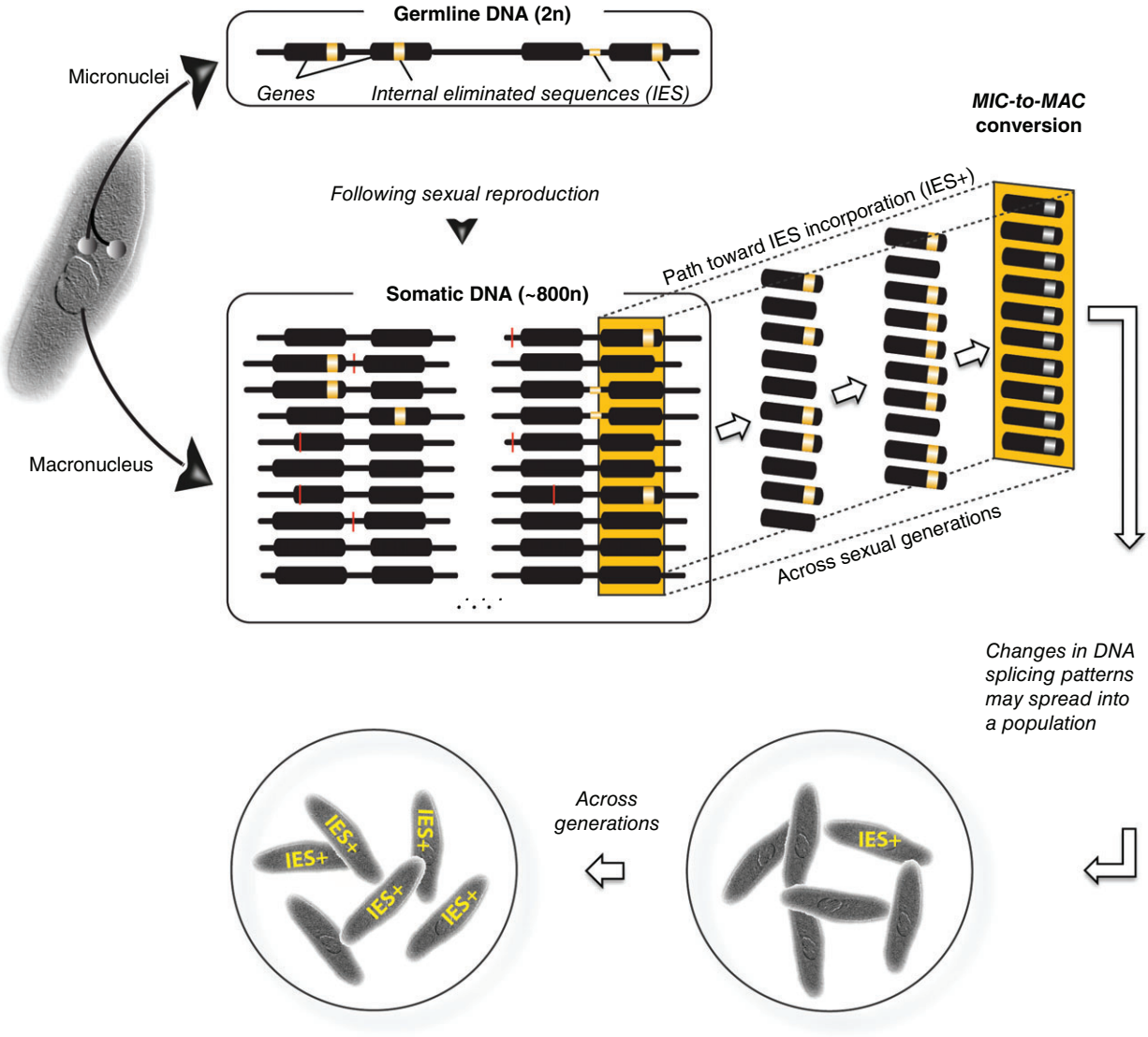


FIGURE 1 | Proposed mechanism for the origin of genetic novelties in Paramecium. In Paramecium programmed DNA elimination occurs when the germline (zygotic) DNA regenerates the polyploid somatic DNA. During this process of germline-to-soma differentiation, the germline DNA is fragmented and amplified and thousands of germline noncoding DNA regions (denoted by the yellow segments in germline DNA) are excised. These excised regions are called internal eliminated sequences (IESs). IESs may be occasionally retained in the somatic DNA (denoted by yellow segments in somatic DNA). Additionally, the DNA splicing machinery may erroneously recognize and excise non-germline specific sequences that are flanked by sequences resembling IES excision signals (as represented by vertical red segments in somatic DNA). Splicing-weakening or -disrupting mutations, in tandem with epigenetic maternal effects, are proposed to modulate the frequency of imperfectly excised sequences. After a number of sexual generations, germline-specific (somatic) sequences can convert into somatic (germline-specific) sequences, a heritable change. Novel DNA variants, when nonlethal, have a nonzero probability of spreading through a population and to reach fixation.

mutations, splicing-weakening mutations might ultimately lead to the complete retention of IESs in the macronucleus and, thus, to the effective conversion of germline-specific sequences into somatic DNA, a process that we term *MIC-to-MAC* conversion or *macronuclearization*. On the other hand, an increased frequency of spliced macronuclear DNA regions might be achieved via mutations strengthening the cryptic excision signals that flank these regions. The optimization of these signals may ultimately facilitate the complete excision of somatic DNA and, thus, the conversion of somatic DNA into germline-specific sequence, a process that we term *MAC-to-MIC* conversion or *micronuclearization*. In all this, it is likely that epigenetic mechanisms contribute to modulating isoform frequencies, regardless of the appearance of new mutations in the *cis*-acting DNA splicing signals.

RNA SPLICING

Eukaryotic protein-coding genes contain intervening sequences, termed spliceosomal introns, which are removed from nascent transcripts through the nuclear process of RNA splicing. Nearly 40 years after their discovery,[33–36] there is greater consensus regarding what introns *are* than why genes have introns. It is commonly thought that introns are neutrally evolving sequences, whose processing is costly.[37] The cost of introns may result from increased time and energy required to transcribe genes.[38,39] That said, studies across a wide range of eukaryotic species have demonstrated that introns can in fact boost gene expression.[40–44] Additionally, or alternatively, the cost of introns may result from potential errors induced by the splicing process.

The origin of spliceosomal introns is equally unclear. A commonly accepted hypothesis maintains that spliceosomal introns originated from a different class of introns that are capable of self-splicing and are frequently found in bacteria, the group II introns.[45,46] Under this hypothesis, group II introns invaded the once intronless nuclear DNA of the eukaryotic ancestor. Following proliferation, intranuclear group II introns lost the ability to self-splice and gave rise to spliceosomal introns and to the machinery that is required for their removal, the spliceosome. Although plausible,[47–50] this hypothesis suffers from at least two important limitations. First, it is not falsifiable. Second, as it stands, it is not clear how this hypothesis can be reconciled with the established fact that RNA splicing is not an isolated process. Indeed, the hypothesis of group II intron origin for spliceosomal introns fails to account for the extensive integration of RNA splicing into the existing network of mRNA-associated processes (see below).

In sum, commonly held views fail to provide definitive answers for the function and the origin(s) of spliceosomal introns. A fresh perspective for understanding these issues may be gained by focusing on the mechanics of RNA splicing.

The Mechanics of RNA Splicing

Several biochemical steps are required for the accurate removal of spliceosomal introns.[51] Here, we place emphasis on two aspects that are important for understanding the following sections. First, introns contain signals that are critical for their excision. These signals are positioned at the intron termini—the donor or 5′ splice site (5′ss), and the acceptor or 3′ splice site (3′ss)—and in the intron body, a branch site, and a polypyrimidine tract. Second, these signal sequences are recognized and bound by the spliceosome. This is a nuclear apparatus that assembles anew at each round of splicing and consists of five U-rich small nuclear (sn)RNAs—U1, U2, U4, U5, and U6—and hundreds of proteins. The association of U1 snRNP with the 5′ss is critical in determining the efficiency of the spliceosome assembly.[52,53] When U1 snRNP binds to the 5′ss, a specific region of the U1 snRNA matches the sequence at and around the 5′ss.[54] If the match between the U1 snRNA (or other RNA-binding snRNAs) and the corresponding splicing signal is optimal, introns are likely to be accurately excised at each round of transcription. Otherwise, introns may be excised in some rounds of transcription and may be retained in others,[55] a process known as alternative RNA splicing. In addition to transcripts that retain introns, alternative RNA splicing may generate other types of isoforms including, for example, transcripts with skipped exons, or transcripts with alternative donor and/or acceptor splice sites.[56] As a result of these many potential alternative rearrangements, alternative RNA splicing can generate vast amounts of mRNA variants, which theoretically translate into vast amounts of protein sequence variation.

How much of this variation is truly functional in the cell remains a matter for investigation. Although the functional role of alternative RNA splicing has been shown in many occasions,[57–60] several studies suggest that in most cases this process simply results from an inaccurate RNA splicing.[61–67] Consistent with this, the alternative RNA splicing-mediated rearrangements of gene sequences often give rise to mRNA variants containing premature translation termination codons (PTCs) or in-frame stops.[68] Although these variants may serve as gene expression regulators,[69,70] they do not lead to protein synthesis in that they are degraded in the cytoplasm by cellular surveillance systems such as the nonsense-mediated decay (NMD) pathway.[71]

RNA Splicing: Its Evolutionary Significance

Studies typically gloss over or underplay the challenges posed by alternative RNA splicing to the traditional definitions of exons and introns.[72] Within protein-coding genes, exons are conventionally understood as DNA regions whose information is incorporated into mature mRNAs. However, exonic information may be left out from mature mRNAs as a result of alternative RNA splicing. Conversely, as a result of alternative RNA splicing, intronic regions may be retained into mature mRNA as if they were exonic. In sum, the classification of intragenic sequences as introns and exons may be fuzzier than typically recognized.

To further complicate matters, studies suggest that introns and exons are less distinct over evolutionary time than conventionally thought. Specifically, it has been shown that a considerable number of exonic and intronic regions result from processes of sequence conversion, which have been labeled exonization (i.e., intronic sequences become exonic)[73,74] and intronization (i.e., exonic sequences become intronic).[75,76] For example, >90% of the ~2700 new exons that have emerged in rodents since the split with humans likely originate from pre-existing unique intronic regions.[77] On the other hand, 16 new introns have emerged in *Caenorhabditis elegans* as a result of the recruitment of internal exonic sequences, an estimate that makes intronization the top contributor to intron creation in this species.[75]

Both exonization and intronization have been observed in several multicellular taxa, including fungi,[78] plants,[79] and vertebrates.[80–83] Of note, sequences undergoing exonization and intronization are alternatively spliced and, in the case of exonization, the ratio of inclusion to exclusion isoforms increases gradually over evolutionary time.[84] Whether intronization involves a similar (though reverse) gradual process of conversion is not yet known.

In conclusion, several studies hint at a mutual conversion process of exonic and intronic sequences over evolutionary time. This implies that a number of alternative RNA splicing events that we observe today could be no more than snapshots of ongoing evolutionary processes, phases that might ultimately lead to new exonic or intronic regions.[76,85] An evaluation of the plausibility and possible ramifications of this hypothetical scenario requires a more extensive understanding of the processes of exonization and intronization.

Exonization

About 150 million years ago in the common therian ancestor of marsupials and placentals, a significant development led to a high frequency of genomic changes that also challenged the splicing system. At that time the Long INterspersed Element (LINE1) retropositional machinery emerged with an exceptionally high activity that continues to the present day. Unlike their precursor forms, LINE1s indiscriminatingly co-retropose any polyadenylated RNA, including the most abundant, polymerase III (pol III)-transcribed Short INterspersed Elements (SINEs). With the random tendency of genomic insertions, SINEs, in particular, integrated into the neighborhood of genes with the potential to change gene expression.

Inserted antisense into existing genes and with reverse oligoA tails, such retroposons introduce a strong polypyrimidine tract (PPT; T_n) splicing motive. Furthermore, in close proximity to the polypyrimidine tract, many SINEs contain one or more cryptic 3′ AG splice sites.[74] With these two splice motives on board, they need only to acquire an additional 5′ GT splice site, which can be part of the element or located somewhere in the flanking intronic sequence, to yield a new exon. If the newly embedded exonized sequence conserves the original open reading frame (ORF) of the subsequent exons, new properties may evolve. If not, the interrupted ORF usually undergoes NMD, or in the worst case leads to a genetic disease, whereby evolution disposes of the destroying form over time.

A newly arisen exon is usually free to be randomly redesigned, possibly to gain new features that might deliver advantages for the individual. The process of exon gain is known as exonization, the acquisition of a new function is exaptation. This 'trial-and-win' game is possible as long as the original function of the host gene is ensured via alternative splicing. As gene evolution progresses, new variants may gain more importance and stronger levels of expression resulting from the evolution of stronger splice sites which allow for increased synthesis of these new, alternative isoforms. In an advanced state, new variants may predominate and replace the original splice variants.[86]

Evolutionary innovation via exonization requires time to be established and to be finally fixed over a few million years in a species. Exonization processes might be rather complex and can transcend the splicing level by including A-to-I RNA editing of the pre-mRNA, leading to novel functional splice sites.[87] Younger exonizations come and go over time unless they acquire some selective advantage to be perpetuated in the transcriptome and possibly proteins of a species. Some examples demonstrate the long and successful evolutionary way of such candidates. An elegant example is the LINE-like, retroposon-derived telomere terminal transferase (telomerase) that functions to elongate from replication to replication-shortened ends of chromosomes in vertebrates.[88] A much younger 'invention' is the endogenous viral elements that encode the provirus Env polyprotein, also known as syncytin, which is important in the architecture of the placenta of many mammalian lineages.[89]

However, uncontrolled transposition would disturb the integrity of the cell and an organism has to find strategies to suppress an unlimited spread of transposable elements (TEs). Cells have invented many regulatory mechanisms to control the

transcription of TEs, fixation, splicing of exonized modules, and the accumulation of harmful exonized mRNAs (see Figure 2). Some of the most important regulatory mechanisms are described hereinafter. (1) Epigenetic silencing of TE transcription occurs when methyl-CpG-binding domains (MBD) attach to methyl groups at CpG dinucleotides and block transcription. This strategy is very effective because TEs harbor more than 50% of all genomic CpG sites.[90] (2) RNA interference destroys, e.g., SINEs when an inversely transcribed TE binds to the TE RNA and, as double-stranded RNA, attracts the RNA-induced silencing complex (RISC) to slice the transcripts.[91] (3) A natural bottleneck for exonized genes is their distribution in the population. Depending on the effective population size, exonized genes need several million years to be fixed in the germline of all members of a species.[92] (4) Direct competitive RNA binding of the heterogeneous nuclear ribonucleoprotein C (HNRNPC—repressor) and the U2 auxiliary factor (U2AF65—specific recognition of cryptic splice sites) regulate alternative splicing.[93] (5) The NMD pathway detects premature termination codons possibly introduced by exonization and triggers the decay of such nonsense mRNAs.[94]

Nevertheless, there is no watertight control over active random insertions of transposed elements. Cell stress (e.g., during viral infections), in particular, leads to increased TE activity and often to disease-causing insertions within functional splice sites or to exonizations that disrupt the ORF of subsequent exons. Our heightened methods of detection have currently identified 40 cases, most of them SINE insertions in antisense orientation, where cryptic TE exons cause disease.[95] Gyrate atrophy of the choroid and retina may lead to blindness and is predetermined by an *Alu*Sg SINE insertion that took place some 35 million years ago in the third intron of the ornithine aminotransferase gene of the common ancestor of Old World monkeys and human primates. Today, a single additional point mutation can convert this insertion to a disease-causing exonization. Many more examples of malfunctioning exonizations are compiled in Vorechovsky 2010. In the end, an ongoing balancing act occurs between innovation and disease-causing exonizations; the latter is the price paid by the try-and-error game of evolution, not stable over evolutionary time, but inevitable and tragic for the carrier.

Intronization

The acquisition of splice sites is critical for the emergence of exons from intronic sequences. The strength of these splice sites (i.e., the efficiency with which these sites are recognized by the spliceosome) will affect the frequency with which new exons are included into mature mRNA isoforms.[70] Moreover, when present in the recruited intronic sequences, in-frame stop codons must be eliminated to guarantee the functionality of the inclusion isoforms. Consistent with this idea, *young* exons, i.e., exons included in minor isoforms, bear more frequently PTCs compared to *older* exons.[96] Similarly, the acquisition of splice sites is critical for the conversion of exonic sequences into intronic regions. Splice sites alone, though, may not guarantee that exonic regions become constitutively spliced introns over evolutionary time. In their *intronization* model,[76] Catania and Lynch propose that the acquisition of PTCs can facilitate constitutive intronization. The proposed mechanism is exemplified hereinafter.

Let us imagine an intronless gene with a PTC that is flanked by cryptic splice sites. During transcription, the spliceosome may occasionally recognize the suboptimal splicing signals, thereby generating two populations of transcripts, i.e., PTC-containing and PTC-free mRNAs. While NMD degrades the PTC-containing mRNAs, the PTC-free mRNAs, which are effectively invisible to NMD, may guarantee the generation of functional proteins with an internal deletion. Except for this deletion, these proteins will be identical to their original (or pre-PTC) version if the spliced region has a size that is multiple of three (or 3*n*) and falls between codons (phase 0). Moreover, these proteins can maintain their original function or acquire a new function provided that the spliced region is sufficiently short and/or nonessential. Thus, under the intronization model, *young* introns are expected to display specific features, e.g., they should be short, 3*n* in length, and contain one or more PTCs. It is worth noting that sequences that are currently categorized as young exons display the same features. The splicing of newborn introns may be initially inefficient because of the weakness of the flanking splice sites. Under these circumstances, splicing may become constitutive if the spliced allele is subjected to positive selection for subsequent mutations that strengthen the splicing signals. Alternatively, splicing of young introns could be efficient from the start if the intronized sequences are located in proximity of splicing-enhancing determinants. One such determinant is the cap-binding complex (CBC), which enhances the association of U1 snRNP with the 5′-most 5′ss of nascent pre-mRNAs.[97–99]

The NMD pathway plays a central role in the intronization hypothesis, a notion consistent with the observation that lineages lacking core NMD components have few introns or none at all.[100] As a

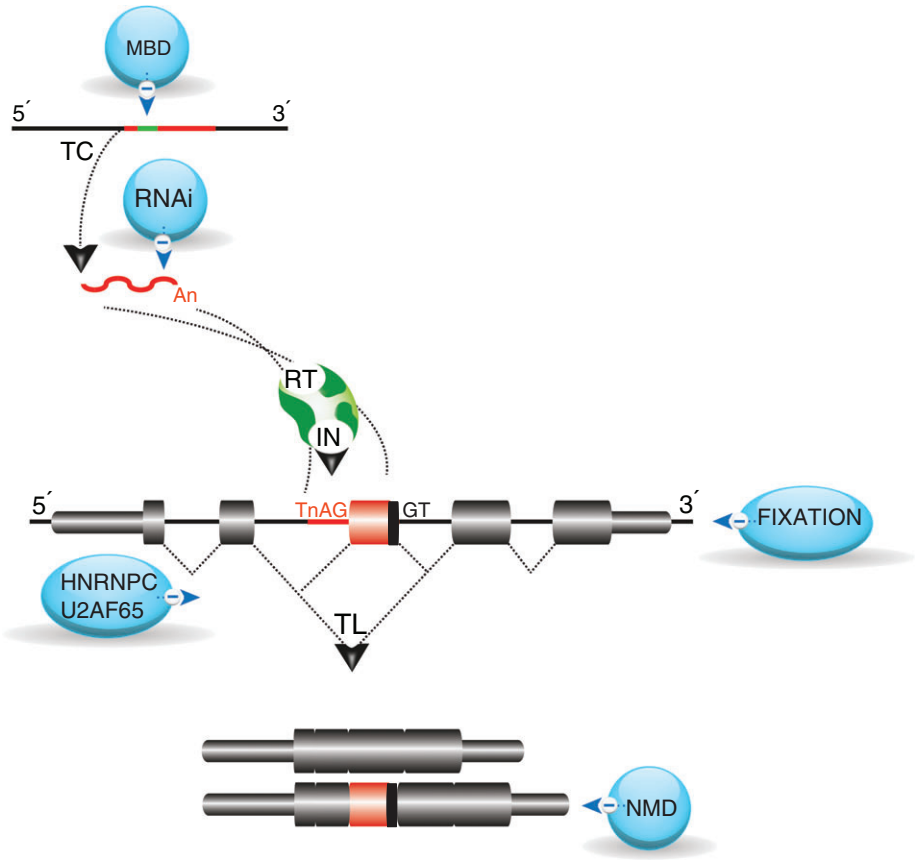


FIGURE 2 | Process of exonization and cellular strategies for diminishing their negative impact. From top to bottom: transposed Short INterspersed Elements (SINEs, red line) are actively transcribed from their own internal polymerase III promoter (green line). Transcription can be downregulated by the methyl-CpG-binding domain (MBD) that binds specifically to methyl CpG pairs. SINE transcripts (TC) are frequently attacked and destroyed by RNA interference (RNAi). Other polyadenylated transcripts are substrates of the reverse transcriptase (RT)- and integrase (IN)-containing retropositional system delivered in trans from active autonomous elements such as LINE1. Random antisense intronic integrations provide an intrinsic polypyrimidine tract (Tn) and frequently also a specific cryptic 3′-AG splice site. Subsequent acquisition of a 5′ GT splice site (in this example located in the flanking intronic region, black box) leads to a composed exonized sequence. To be stably inherited in a species any genomic innovation requires germline fixation, a process that takes millions of years. Alternative splicing of an exonized sequence is regulated by the competitive activity of the heterogeneous nuclear ribonucleoprotein C (HNRNPC) that functions as a repressor and the U2 auxiliary factor (U2AF65) that actively processes cryptic splice sites occurring in newly exonized sequences. Nonfunctional mRNAs with premature termination codons introduced by exonized sequences (red and black boxes) will be recognized and destroyed in the nonsense-mediated decay (NMD) pathway during translation (TL).

result, it is likely that NMD properties have significant repercussions on the process of intronization. A commonly acknowledged property of NMD—the NMD-mediated degradation of aberrant transcripts is least efficient when in-frame stops are close to the mRNA tail[101,102]—may affect the probability of intronization along the intragenic territory (see Figure 3). This latter observation, together with the splicing-enhancing role of the CBC at the pre-mRNA 5′ end, points to a directional effect. Namely, not only is the pool of spliced variants relatively pure and associated with minimal fitness effects when PTCs are located at the gene 5′ end, but splicing at this location might also be facilitated by the CBC. This extended conclusion gives further support to the idea that intronization is more likely to occur at the gene 5′ end rather than the gene 3′ end. If we assume a steady-state process of intron birth and death,[37] this implies that the spatial distribution of introns within genes should be biased toward the 5′ end. Notably, a 5′ end positional bias of introns is detected across multiple eukaryotes.[103]

ALTERNATIVE SPLICING AND ITS ROLE IN SEQUENCE CONVERSION

That exonic and intronic sequences may convert into one another clashes with the classical view that these sequences evolve as separate, watertight, compartments. Intriguingly, the proposed processes of

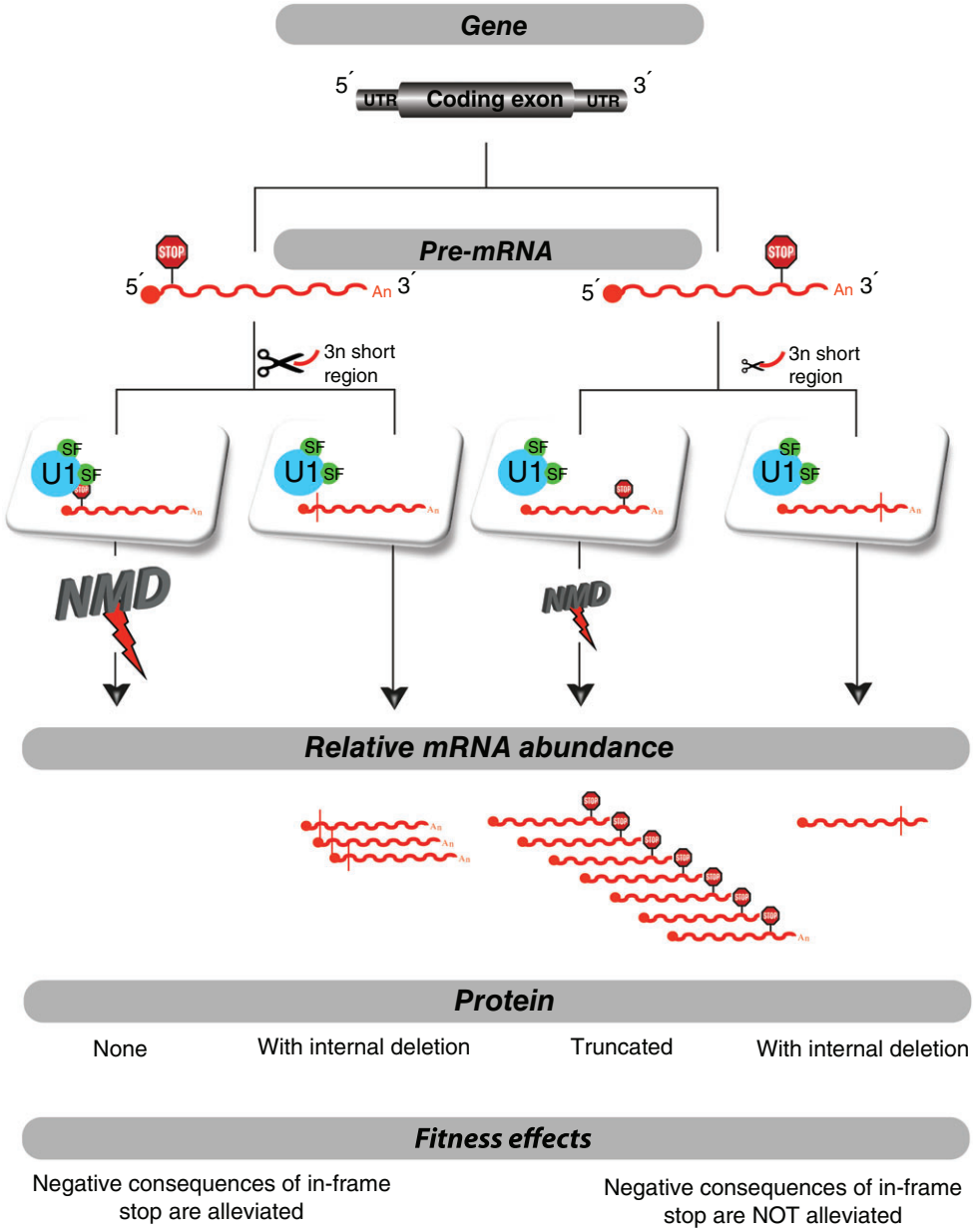


FIGURE 3 | Overview of the proposed mechanism of intronization. A gene acquires a premature in-frame stop codon either at its 5′ end or at its 3′ end. This stop codon may or may not be removed from the corresponding pre-mRNA via RNA splicing provided that it is flanked by latent splicing signals. If accidental splicing takes place, and the excised sequence falls between codons and contains a number of nucleotides that is a multiple of 3 (3*n*), then the resulting mature mRNA will be invisible to nonsense-mediated decay (NMD) and the translation product will be identical to the original protein, except for an internal deletion. Accidental splicing is more likely to take place in proximity of pre-mRNA structures/sequences that enhance the recruitment of splicing factors on site, e.g., the cap-binding complex at the gene 5′ end. If splicing does not take place, NMD will produce a relatively pure pool of stop-free mRNAs when the stop codon resides at the gene 5′ end—at this location NMD degradation of aberrant transcripts is most efficient. In contrast, truncated and potentially harmful translational products are produced when mRNAs contain premature in-frame stops at their 3′ end. As a consequence of the expected fitness effects that are associated with these dynamics, intronization is more likely to occur at the gene 5′ end compared to the gene 3′ end.

sequence conversion, which entail gene regions that are neither fully coding nor fully noncoding, may be unfolding under our own eyes in the form of sequences undergoing alternative splicing in modern species. As discussed above, there exist two major views on the biological role of alternative RNA splicing, one where this process contributes to organismal complexity by generating regulatory and protein diversity, and one

where alternative variants are mostly noise, the mere result of suboptimal signal sequences. Each of these views is supported by several findings, thus a legitimate question is: 'How can the hypothesized role of alternative RNA splicing in facilitating sequence conversion be integrated with these views?'

While one may set the beginning of a process of sequence conversion to the time when the splicing machineries recognize suboptimal splicing signals, it is less obvious to predict how splicing events could unfold. Because the levels (and the patterns) of alternative RNA splicing are inevitably contingent on stochastic mutations altering the quality of the splicing signals, newly generated nonfunctional isoforms could become increasingly frequent over evolutionary time and facilitate the discussed events of sequence conversion or remain at low frequency or disappear. Moreover, at any point in time alternative isoforms may acquire functional relevance, a condition that facilitates the preservation of the alternative RNA splicing state. Hence, a scenario in which alternative RNA splicing facilitates the mutual conversion between exonic and intronic sequences is compatible with current findings supporting the alternative splicing-mediated generation of both functional and nonfunctional RNA splicing diversity. In the same way, alternative DNA splicing facilitating the mutual conversion between somatic and germline (coding/noncoding) sequences[30] is compatible with the occurrence of nonfunctional[9,16] and functional[104] alternative DNA-level splicing.

Cell Biology Meets Population Genetics

Natural selection shapes the quality of splicing signals as well as some of the properties of the DNA and RNA splicing substrates. For example, a significant deficit of 3*n* and PTC-free short introns and exon-mapping IESs has been detected in several eukaryotes and in Paramecium, respectively.[16,105] This deficit presumably results from the hazard associated with these sequences, which, if retained, would be invisible to NMD and would generate potentially toxic proteins. Also, both IESs and introns that are non-3*n* and bear in-frame stops are located preferentially at the gene 5′ end in Paramecium.[30,76] As described above, this positional bias optimizes the NMD efficiency of recognition/degradation of PTC-containing alleles. Finally, a significant deficit of cryptic DNA and RNA splicing signals has been detected in a number of eukaryotes, consistent with the idea that these signals are counter-selected to prevent excisions that may have deleterious effects on fitness.[32,106]

By regulating the strength of splicing signals, natural selection should play a key role in shaping isoform frequencies over evolutionary time. Provided that splicing-weakening mutations that generate alternative isoforms are often slightly deleterious, they should accumulate preferentially in species with a small enough effective population size that the power of random genetic drift is in excess of the power of selection. Although this expectation has yet to be verified in the case of programmed DNA elimination, in the case of RNA splicing current observations support this scenario: as the power of natural selection decreases with increasing organism size,[107] the fraction of alternatively RNA splicing increases from unicellular to multicellular eukaryotic species.[108,109] This observation implies that it should be easier to detect *ongoing* processes of sequence conversion in multicellular species than in unicellular species. In a population-genetic environment where selection is efficient, a more rapid establishment of mutations, which accelerate the completion of the transient process of alternative splicing, is expected. It follows that while displaying fewer ongoing events of sequence conversion, unicellular species may exhibit an excess of functional *completed* events of sequence conversion compared to multicellular species.

CONCLUSION

Splicing of nucleic acids contributes to fundamental biological events, including antibody formation,[5] change in the developmental fate of the affected cells,[109] mating types and sex determination,[104,110] and speciation.[111] Here, we have presented several notions consistent with a model where the genetic and epigenetic mechanisms mediating programmed DNA elimination and RNA splicing facilitate an additional fundamental biological event: the mutual conversion between intronic and exonic sequences or between germline and somatic (coding/noncoding) sequences.

Noncoding RNAs are central to guide programmed DNA elimination and RNA splicing. These RNAs facilitate the recognition of excision signal sequences, whose composition is, in turn, shaped by the interplay between evolutionary forces. By means of this interplay, the frequency of alternative isoforms may change so as to facilitate the discussed events of sequence conversion. In essence, the progressive weakening of the IES excision signals may facilitate the permanent incorporation of noncoding IESs into somatic coding DNA sequences (macronuclearization). Moreover, spliced somatic coding and noncoding regions might effectively convert into noncoding germline-specific sequences (micronuclearization) via the optimization of originally cryptic excision signals. Similarly, as a result of fortuitous changes in

the signals specifying RNA splicing, intronic regions may ultimately convert into exons (exonization) and, vice versa, noncoding introns may emerge from exonic sequences (intronization).

In addition to changes in the quality of the splicing signals, the gain (loss) of in-frame stops by spliced coding (noncoding) regions is expected to facilitate sequence conversion in the case of intronization or exonization and also in the case of micronuclearization and macronuclearization, particularly when IESs or spliced somatic DNA regions reside within coding exons. Furthermore, the crosstalk between genetic and epigenetic signals may affect the process of sequence conversion, as well as favor the preservation of the alternative state, if either of the resultant splicing variants is functional.

The compendium of observations and molecular interactions that we provide here, although incomplete, suggests that while a reductionist approach is needed to shed light on crucial mechanistic details of RNA- and DNA-level splicing, a holistic interdisciplinary approach is required to achieve a thorough understanding of the evolutionary significance of the processes of programmed DNA elimination and RNA splicing.

ACKNOWLEDGMENTS

The authors thank Diana Ferro for comments on the first draft of this manuscript. This work was supported by the Deutsche Forschungsgemeinschaft grants to FC (CA1416/1-1) and to JS (SCHM1469/3-2).

REFERENCES

1. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP. Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci USA* 2012, 109:18018–18023.
2. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 2007, 41:331–368.
3. Zufall RA, Robinson T, Katz LA. Evolution of developmentally regulated genome rearrangements in eukaryotes. *J Exp Zool B Mol Dev Evol* 2005, 304:448–455.
4. Kloc M, Zagrodzinska B. Chromatin elimination—an oddity or a common mechanism in differentiation and development? *Differentiation* 2001, 68:84–91.
5. Schatz DG, Ji Y. Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol* 2011, 11:251–263.
6. Klobutcher LA, Herrick G. Developmental genome reorganization in ciliated protozoa: the transposon link. *Prog Nucleic Acid Res Mol Biol* 1997, 56:1–62.
7. Marmignon A, Bischerour J, Silve A, Fojcik C, Dubois E, Arnaiz O, Kapusta A, Malinsky S, Betermier M. Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *PLoS Genet* 2014, 10:e1004552.
8. Lhuillier-Akakpo M, Frapporti A, Denby Wilkes C, Matelot M, Vervoort M, Sperling L, Duharcourt S. Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. *PLoS Genet* 2014, 10:e1004665.
9. Sandoval PY, Swart EC, Arambasic M, Nowacki M. Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting. *Dev Cell* 2014, 28:174–188.
10. Vogt A, Mochizuki K. A domesticated PiggyBac transposase interacts with heterochromatin and catalyzes reproducible DNA elimination in Tetrahymena. *PLoS Genet* 2013, 9:e1004032.
11. Yao MC, Choi J, Yokoyama S, Austerberry CF, Yao CH. DNA elimination in Tetrahymena: a developmental process involving extensive breakage and rejoining of DNA at defined sites. *Cell* 1984, 36:433–440.
12. Mochizuki K, Gorovsky MA. A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes Dev* 2005, 19:77–89.
13. Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, Perlman DH, Doak TG, Stuart A, Amemiya CT, et al. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell* 2014, 158:1187–1198.
14. Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, Landweber LF. RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* 2008, 451:153–158.
15. Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF. Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement. *Cell* 2012, 151:1243–1255.
16. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury JM, Denby Wilkes C, Garnier O, Labadie K, Lauderdale BE, Le Mouel A, et al. The Paramecium germline

genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet* 2012, 8:e1002984.

17. Klobutcher LA, Herrick G. Consensus inverted terminal repeat sequence of Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons. *Nucleic Acids Res* 1995, 23:2006–2013.
18. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, Betermier M. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev* 2009, 23:2478–2483.
19. Dubois E, Bischerour J, Marmignon A, Mathy N, Regnier V, Betermier M. Transposon invasion of the Paramecium germline genome countered by a domesticated PiggyBac transposase and the NHEJ pathway. *Int J Evol Biol* 2012, 2012:436196.
20. Mayer KM, Forney JD. A mutation in the flanking 5'-TA-3' dinucleotide prevents excision of an internal eliminated sequence from the *Paramecium tetraurelia* genome. *Genetics* 1999, 151:597–604.
21. Matsuda A, Mayer KM, Forney JD. Identification of single nucleotide mutations that prevent developmentally programmed DNA elimination in *Paramecium tetraurelia*. *J Eukaryot Microbiol* 2004, 51:664–669.
22. Mayer KM, Mikami K, Forney JD. A mutation in *Paramecium tetraurelia* reveals functional and structural features of developmentally excised DNA elements. *Genetics* 1998, 148:139–149.
23. Duharcourt S, Keller AM, Meyer E. Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in *Paramecium tetraurelia*. *Mol Cell Biol* 1998, 18:7075–7085.
24. Lepere G, Nowacki M, Serrano V, Gout JF, Guglielmi G, Duharcourt S, Meyer E. Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. *Nucleic Acids Res* 2009, 37:903–915.
25. Bouhouche K, Gout JF, Kapusta A, Betermier M, Meyer E. Functional specialization of Piwi proteins in *Paramecium tetraurelia* from post-transcriptional gene silencing to genome remodelling. *Nucleic Acids Res* 2011, 39:4249–4264.
26. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 2006, 444:171–178.
27. Betermier M, Duharcourt S, Seitz H, Meyer E. Timing of developmentally programmed excision and circularization of Paramecium internal eliminated sequences. *Mol Cell Biol* 2000, 20:1553–1561.
28. Betermier M. Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. *Res Microbiol* 2004, 155:399–408.
29. Duret L, Cohen J, Jubin C, Dessen P, Gout JF, Mousset S, Aury JM, Jaillon O, Noel B, Arnaiz O, et al. Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline. *Genome Res* 2008, 18:585–596.
30. Catania F, McGrath CL, Doak TG, Lynch M. Spliced DNA sequences in the Paramecium germline: their properties and evolutionary potential. *Genome Biol Evol* 2013, 5:1200–1211.
31. Duharcourt S, Butler A, Meyer E. Epigenetic self-regulation of developmental excision of an internal eliminated sequence on *Paramecium tetraurelia*. *Genes Dev* 1995, 9:2065–2077.
32. Swart EC, Wilkes CD, Sandoval PY, Arambasic M, Sperling L, Nowacki M. Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Res* 2014, 42:8970–8983.
33. Berget SM, Moore C, Sharp PA. Spliced segments at 5' terminus of adenovirus 2 late messenger-RNA. *Proc Natl Acad Sci USA* 1977, 74:3171–3175.
34. Chow LT, Gelinas RE, Broker TR, Roberts RJ. Amazing sequence arrangement at 5' ends of adenovirus-2 messenger-RNA. *Cell* 1977, 12:1–8.
35. Evans RM, Fraser N, Ziff E, Weber J, Wilson M, Darnell JE. Initiation sites for RNA-transcription in Ad2 DNA. *Cell* 1977, 12:733–740.
36. Goldberg S, Schwartz H, Darnell JE. Evidence from UV transcription mapping in Hela-cells that heterogeneous nuclear-RNA is messenger-RNA precursor. *Proc Natl Acad Sci USA* 1977, 74:4520–4523.
37. Lynch M. Intron evolution as a population-genetic process. *Proc Natl Acad Sci USA* 2002, 99:6118–6123.
38. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. Selection for short introns in highly expressed genes. *Nat Genet* 2002, 31:415–418.
39. Seoighe C, Gehring C, Hurst LD. Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genet* 2005, 1:e13.
40. Rose AB. The effect of intron location on intron-mediated enhancement of gene expression in Arabidopsis. *Plant J* 2004, 40:744–751.
41. Buchman AR, Berg P. Comparison of intron-dependent and intron-independent gene-expression. *Mol Cell Biol* 1988, 8:4395–4405.
42. Callis J, Fromm M, Walbot V. Introns increase gene expression in cultured maize cells. *Genes Dev* 1987, 1:1183–1200.
43. Duncker BP, Davies PL, Walker VK. Introns boost transgene expression in *Drosophila melanogaster*. *Mol Gen Genet* 1997, 254:291–296.
44. Palmiter RD, Sandgren EP, Avarbock MR, Allen DD, Brinster RL. Heterologous introns can enhance expression of transgenes in mice. *Proc Natl Acad Sci USA* 1991, 88:478–482.

45. Cavalier-Smith T. Selfish DNA and the origin of introns. *Nature* 1985, 315:283–284.
46. Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* 2006, 1:22.
47. Cavalier-Smith T. Intron phylogeny—a new hypothesis. *Trends Genet* 1991, 7:145–148.
48. Jacquier A. Self-splicing group-II and nuclear pre-messenger-RNA introns—how similar are they. *Trends Biochem Sci* 1990, 15:351–354.
49. Cousineau B, Lawrence S, Smith D, Belfort M. Retrotransposition of a bacterial group II intron. *Nature* 2000, 404:1018–1021.
50. Palmer JD, Logsdon JM Jr. The recent origins of introns. *Curr Opin Genet Dev* 1991, 1:470–477.
51. Will CL, Luhrmann R. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* 2011, 3: a003707.
52. Seraphin B, Rosbash M. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. *Cell* 1989, 59:349–358.
53. Michaud S, Reed R. An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes Dev* 1991, 5:2534–2546.
54. Kondo Y, Oubridge C, van Roon AM, Nagai K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife* 2015, 4:e04986.
55. Kuo HC, Nasim FH, Grabowski PJ. Control of alternative splicing by the differential binding of U1 small nuclear ribonucleoprotein particle. *Science* 1991, 251:1045–1050.
56. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 2008, 14:802–813.
57. Bell LR, Horabin JI, Schedl P, Cline TW. Positive autoregulation of sex-lethal by alternative splicing maintains the female determined state in Drosophila. *Cell* 1991, 65:229–239.
58. Izquierdo JM, Majos N, Bonnal S, Martinez C, Castelo R, Guigo R, Bilbao D, Valcarcel J. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell* 2005, 19:475–484.
59. Rosenfeld MG, Lin CR, Amara SG, Stolarsky L, Roos BA, Ong ES, Evans RM. Calcitonin mRNA polymorphism: peptide switching associated with alternative RNA splicing events. *Proc Natl Acad Sci USA* 1982, 79:1717–1721.
60. Chen L, Kostadima M, Martens JH, Canu G, Garcia SP, Turro E, Downes K, Macaulay IC, Bielczyk-Maczynska E, Coe S, et al. Transcriptional diversity during lineage commitment of human blood progenitors. *Science* 2014, 345:1251033.
61. Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet* 2003, 12:1313–1320.
62. Uyar B, Chu JS, Vergara IA, Chua SY, Jones MR, Wong T, Baillie DL, Chen N. RNA-seq analysis of the *C. briggsae* transcriptome. *Genome Res* 2012, 22:1567–1580.
63. Nurtdinov RN, Neverov AD, Favorov AV, Mironov AA, Gelfand MS. Conserved and species-specific alternative splicing in mammalian genomes. *BMC Evol Biol* 2007, 7:249.
64. Malko DB, Makeev VJ, Mironov AA, Gelfand MS. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome Res* 2006, 16:505–509.
65. Wang BB, Brendel V. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci USA* 2006, 103:7175–7180.
66. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* 2010, 6:e1001236.
67. Wang M, Zhang P, Shu Y, Yuan F, Zhang Y, Zhou Y, Jiang M, Zhu Y, Hu L, Kong X, et al. Alternative splicing at GYNNGY 5' splice sites: more noise, less regulation. *Nucleic Acids Res* 2014, 42:13969–13980.
68. Sorek R, Shamir R, Ast G. How prevalent is functional alternative splicing in the human genome? *Trends Genet* 2004, 20:68–71.
69. 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.
70. Baek D, Green P. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci USA* 2005, 102:12813–12818.
71. Chang YF, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* 2007, 76:51–74.
72. Gilbert W. Why genes in pieces. *Nature* 1978, 271:501.
73. Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, Graur D, Ast G. Minimal conditions for exonization of intronic sequences: 5' splice site formation in Alu exons. *Mol Cell* 2004, 14:221–231.
74. Lev-Maor G, Sorek R, Shomron N, Ast G. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* 2003, 300:1288–1291.
75. Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW. Origin of introns by 'intronization' of exonic sequences. *Trends Genet* 2008, 24:378–381.
76. Catania F, Lynch M. Where do introns come from? *PLoS Biol* 2008, 6:e283. doi:10.1371/journal.pbio.0060283.

77. Wang W, Zheng HK, Yang S, Yu HJ, Li J, Jiang HF, Su JN, Yang L, Zhang JG, McDermott J, et al. Origin and evolution of new exons in rodents. *Genome Res* 2005, 15:1258–1264.

78. Roy SW. Intronization, de-intronization and intron sliding are rare in Cryptococcus. *BMC Evol Biol* 2009, 9:192.

79. Zhu Z, Zhang Y, Long M. Extensive structural renovation of retrogenes in the evolution of the Populus genome. *Plant Physiol* 2009, 151:1943–1951.

80. Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I. Primate and rodent specific intron gains and the origin of retrogenes with splice variants. *Mol Biol Evol* 2011, 28:33–37.

81. Kang L, Zhu Z, Zhao Q, Chen L, Zhang Z. Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. *BMC Evol Biol* 2012, 12:128.

82. Zhang XHF, Chasin LA. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci USA* 2006, 103:13427–13432.

83. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* 2014, 24:1774–1786.

84. Alekseyenko AV, Kim N, Lee CJ. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA* 2007, 13:661–670.

85. Catania F, Gao X, Scofield DG. Endogenous mechanisms for the origins of spliceosomal introns. *J Hered* 2009, 100:591–596.

86. Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J. Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res* 2007, 17:1139–1145.

87. Möller-Krull M, Zemann A, Roos C, Brosius J, Schmitz J. Beyond DNA: RNA editing and steps toward Alu exonization in primates. *J Mol Biol* 2008, 382:601–609.

88. Bowen NJ, Jordan IK. Exaptation of protein coding sequences from transposable elements. *Genome Dyn* 2007, 3:147–162.

89. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta* 2012, 33:663–671.

90. Xie H, Wang M, Bonaldo Mde F, Rajaram V, Stellpflug W, Smith C, Arndt K, Goldman S, Tomita T, Soares MB. Epigenomic analysis of Alu repeats in human ependymomas. *Proc Natl Acad Sci USA* 2010, 107:6952–6957.

91. Soifer HS, Zaragoza A, Peyvan M, Behlke MA, Rossi JJ. A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. *Nucleic Acids Res* 2005, 33:846–856.

92. Schmitz J, Zischler H. Molecular cladistic markers and the infraordinal phylogenetic relationships of primates. In: Kay RF, Ross C, eds. *Anthropoid Origins: New Visions*. New York: Kluwer Academic Press; 2004, 57–69.

93. Zarnack K, Konig J, Tajnik M, Martincorena I, Eustermann S, Stevant I, Reyes A, Anders S, Luscombe NM, Ule J. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* 2013, 152:453–466.

94. Sela N, Mersch B, Hotz-Wagenblatt A, Ast G. Characteristics of transposable element exonization within human and mouse. *PLoS One* 2010, 5:e10907.

95. Vorechovsky I. Transposable elements in disease-associated cryptic exons. *Hum Genet* 2010, 127:135–154.

96. Xing Y, Lee CJ. Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy. *Trends Genet* 2004, 20:472–475.

97. Lewis JD, Izaurralde E, Jarmolowski A, McGuigan C, Mattaj IW. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev* 1996, 10:1683–1698.

98. Inoue K, Ohno M, Sakamoto H, Shimura Y. Effect of the cap structure on pre-mRNA splicing in Xenopus oocyte nuclei. *Genes Dev* 1989, 3:1472–1479.

99. Ohno M, Sakamoto H, Shimura Y. Preferential excision of the 5' proximal intron from mRNA precursors with two introns as mediated by the cap structure. *Proc Natl Acad Sci USA* 1987, 84:5187–5191.

100. Lynch M, Hong X, Scofield DG. NMD and the evolution of eukaryotyic gene structure. In: Maquat LE, ed. *Nonsense-Mediated mRNA Decay*. Georgetown, TX: Landes Bioscience; 2006, 197–211.

101. van Hoof A, Green PJ. Premature nonsense codons decrease the stability of phytohemagglutinin mRNA in a position-dependent manner. *Plant J* 1996, 10:415–424.

102. Longman D, Plasterk RH, Johnstone IL, Caceres JF. Mechanistic insights and identification of two novel factors in the *C. elegans* NMD pathway. *Genes Dev* 2007, 21:1075–1085.

103. Lin K, Zhang DY. The excess of 5' introns in eukaryotic genomes. *Nucleic Acids Res* 2005, 33:6522–6527.

104. Singh DP, Saudemont B, Guglielmi G, Arnaiz O, Gout JF, Prajer M, Potekhin A, Przybos E, Aubusson-Fleury A, Bhullar S, et al. Genome-defence small RNAs exapted for epigenetic mating-type inheritance. *Nature* 2014, 509:447–452.

105. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Segurens B, et al. Translational control of intron splicing in eukaryotes. *Nature* 2008, 451:359–362.

106. Farlow A, Dolezal M, Hua L, Schlotterer C. The genomic signature of splicing-coupled selection differs between long and short introns. *Mol Biol Evol* 2012, 29:21–24.

107. Lynch M. *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates; 2007.

108. Catania F, Lynch M. A simple model to explain evolutionary trends of eukaryotic gene architecture and expression: how competition between splicing and cleavage/polyadenylation factors may affect gene expression and splice-site recognition in eukaryotes. *Bioessays* 2013, 35:561–570.

109. Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. *Mol Biol Evol* 2014, 31:1402–1413.

110. Valcarcel J, Singh R, Zamore PD, Green MR. The protein sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of transformer pre-mRNA. *Nature* 1993, 362:171–175.

111. Terai Y, Morikawa N, Kawakami K, Okada N. The complexity of alternative splicing of hagoromo mRNAs is increased in an explosively speciated lineage in East African cichlids. *Proc Natl Acad Sci USA* 2003, 100:12798–12803.

