Published online 24 August 2015

Nucleic Acids Research, 2015, Vol. 43, No. 17 ***8157–8168***
doi: 10.1093/nar/gkv843

Cis-acting signals modulate the efficiency of programmed DNA elimination in *Paramecium tetraurelia*

Diana Ferro†, Gildas Lepennetier† and Francesco Catania*

Institute for Evolution and Biodiversity, University of Münster, Hüfferstrasse 1, 48149 Münster, Germany

Received May 06, 2015; Revised July 07, 2015; Accepted August 01, 2015

ABSTRACT

In Paramecium, the regeneration of a functional somatic genome at each sexual event relies on the elimination of thousands of germline DNA sequences, known as Internal Eliminated Sequences (IESs), from the zygotic nuclear DNA. Here, we provide evidence that IESs' length and sub-terminal bases jointly modulate IES excision by affecting DNA conformation in *P. tetraurelia*. Our study reveals an excess of complementary base pairing between IESs' sub-terminal and contiguous sites, suggesting that IESs may form DNA loops prior to cleavage. The degree of complementary base pairing between IESs' sub-terminal sites (termed C_{in}-score) is positively associated with IES length and is shaped by natural selection. Moreover, it escalates abruptly when IES length exceeds 45 nucleotides (nt), indicating that only sufficiently large IESs may form loops. Finally, we find that IESs smaller than 46 nt are favored targets of the cellular surveillance systems, presumably because of their relatively inefficient excision. Our findings extend the repertoire of *cis*-acting determinants for IES recognition/excision and provide unprecedented insights into the distinct selective pressures that operate on IESs and somatic DNA regions. This information potentially moves current models of IES evolution and of mechanisms of IES recognition/excision forward.

INTRODUCTION

Single-celled ciliated protozoa are excellent systems for the study of programmed DNA elimination. These protozoa are the only eukaryotes with two nuclei in their cytoplasm: a micronucleus (MIC), which houses the germline genome, and a macronucleus (MAC) that contains the somatic genome. In ciliates, programmed DNA elimination takes place in the new zygotic MAC—as the old maternal MAC degrades—and regulates the excision of up to tens of thousands of germline DNA sequences, known as Internal Eliminated Sequences (or IESs). Programmed DNA elimination (also referred to as DNA splicing hereinafter) in ciliates bears upon the generation of a functional somatic genome and the survival of sexual progeny.

Paramecium is one of the best-studied genera of ciliates, and *P. tetraurelia* is its best-characterized species (1). *P. tetraurelia*'s somatic genome is small (72 Mb), compact (78% coding density), AT-rich (72%), gene-rich (∼40 000 genes) and polyploid (∼800n) (2). In addition, the germline DNA of *P. tetraurelia* has been sequenced, revealing the identity and the properties of ∼45 000 IESs. These IESs are short—more than 90% are shorter than 150 base pairs (bp)—and typically single-copy (3).

The molecular mechanisms of programmed DNA elimination in *P. tetraurelia* have been extensively studied. It is clear that DNA splicing in *P. tetraurelia* takes place while the developing MAC is being amplified to reach ∼800 copies (4). During this operation, IESs may be left uncut accidentally at several loci in a variable fraction of macronuclear genome copies (5,6). Furthermore, DNA splicing involves the introduction of DNA double-strand breaks by means of a domesticated transposase, PiggyMac (7,8), as well as the deposition of specific histone marks (9), and the participation of a number of molecular actors (10–12). Prominent among these actors are two distinct classes of small RNAs: scnRNAs and iesRNAs. scnRNAs assist in the excision of developing MAC sequences that are absent from the maternal (pre-zygotic) MAC (13–15). This homology-dependent DNA splicing is thought to involve 1/3 or less of *P. tetraurelia* IESs, which are known as maternally-controlled IESs (9). A second class of small RNAs, iesRNAs, facilitate the excision of a fraction of non-maternally controlled IESs (16).

In addition, several observations suggest that a number of genetic features of IESs importantly affect DNA splicing in *Paramecium*. One of these features is the composition of the IES sub-terminal sequences. *P. tetraurelia* IESs are flanked by two 5′-TA-3′ dinucleotides, one of which is

*To whom correspondence should be addressed. Tel: +49 251 83 21222; Fax: +49 251 83 24668; Email: francesco.catania@uni-muenster.de
†These authors contributed equally to the paper as first authors.

retained in the somatic genome after IES excision. These dinucleotides, in turn, are part of 8-bp imperfect inverted terminal repeats (4,17). It has been shown that point mutations that abolish the 5′-TA-3′ dinucleotide (18,19), or alter the termini of a maternally controlled IES (20) cause IES retention. Not only does this imply that IES termini are critical for the process of IES recognition/excision, it also suggests that nucleotide changes in IES termini can override epigenetic regulation.

Another variable that critically affects DNA splicing is IES length. As IES size increases, scnRNAs (but not iesRNAs) become increasingly important for accurate IES excision (16). This implies that the excision of small IESs is generally less sensitive to scnRNAs. In addition, the frequencies of IES sub-terminal bases change with IES size (21). Finally, the mechanism of IES excision itself might be IES length-dependent. Specifically, it has been proposed that IESs might form a double-stranded DNA loop during assembly of an active excision complex (19), provided that IESs are sufficiently large (3). The extent to which DNA conformation affects the recognition/excision of *P. tetraurelia* IESs remains unclear, however, despite the potential insights that it may provide to the mechanisms of IES recognition/excision. For example, if large IESs truly form a loop prior to cleavage, then the observed involvement of scnRNAs in the excision of large (but not small) IESs might be perhaps required to facilitate the crosstalk between IESs' ends. Additionally, the ability to form loops could provide insights into IES evolution. For example, the excision of large (loop-forming) IESs might be inherently less efficient/reliable compared to that of small IESs. This could explain at least in part the great abundance of small IESs in the Paramecium germline genome, and the preferential loss of large IESs over evolutionary time (3).

Here, we asked three questions. First, what is the likelihood that *P. tetraurelia* IESs form loops? We addressed this issue by studying the extent to which the complementary base pairing between (i) intra-IES termini, and (ii) somatic DNA sequences abutting *P. tetraurelia* IESs deviates from what we would expect if pairing occurred at random. Second, we looked into whether and how DNA splicing efficiency is affected by variations in complementary base pairing between IESs' sub-terminal or contiguous sequences. To this end, we compared the IES-associated levels of complementary base pairing in genomic regions that evolve under distinct levels of selective pressure. And third, we leveraged the idea that cellular surveillance systems (such as the Nonsense Mediated Decay (NMD) pathway) should be able to detect recurrent inefficient IES excision, to examine the reliability of mechanisms guiding the excision of large (putatively loop-forming) IESs compared to those guiding the excision of small IESs.

Our results lend support to the proposition that sufficiently large (>45 nucleotides (nt)) IESs may indeed form loops prior to their excision, and that complementary base pairing between IESs' sub-terminal and contiguous sequences may impact IES recognition/excision. The unprecedented insights into the evolution of germline and somatic DNA sequences uncovered by our study sustain and help extend current models of the mechanisms that guide DNA splicing in *P. tetraurelia*.

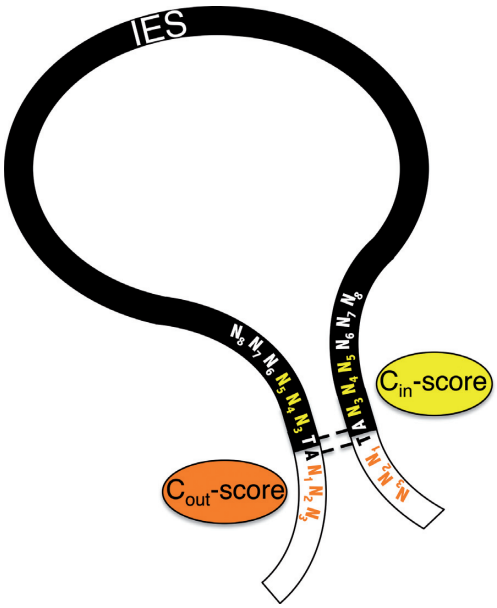

Figure 1. Model of IES loop. Germline and somatic DNA sequences are depicted in black and white, respectively. 'N' indicates any of the four bases (A, C, G, T). The fraction of complementary bases at IES sub-terminal positions 3, 4 and 5 (C_{in}-score), and at IES contiguous positions ±1, ±2 and ±3 (C_{out}-score) is examined as a possible quality measure for *cis*-acting IES recognition/excision signals.

MATERIALS AND METHODS

Surveyed IESs

We examined 43 086 *P. tetraurelia* IESs (35 075 intragenic and 8011 intergenic). Intragenic IESs reside in 17 572 genes (GAZE models), which all begin with the sequence 'ATG'. The surveyed IESs were extracted from files stored in the Paramecium genome database at the URL: http://paramecium.cgm.cnrs-gif.fr/; file names: 'ptetraurelia_CDS_v1_pt_51.gff3.gz', 'parameciumDB.gff3.gz', 'Ptetraurelia_genes_cur.fasta', and 'Ptetraurelia_genes_with_IES_cur.fasta' (22,23). We used in-house bash and python scripts to extract several IES-associated features, including size, relative position along genes, presence/absence of in-frame TGAs, sub-terminal and contiguous sequences, and fraction of complementary base pairing between intra-IES termini positions 3, 4 and 5 (referred to as C_{in}-score; Figure 1) and 6, 7, 8, as well as between IES-flanking positions ±1, ±2, ±3 (referred to as C_{out}-score; Figure 1) and ±4, ±5, ±6.

Gene expression

We investigated the relationship between the IES properties described above and the expression of the IES-mapping genes. Gene expression data, which were previously generated for *P. tetraurelia* (24), were downloaded from Gene Expression Omnibus (25) (see accession number GPL7221). As in Arnaiz *et al.* (3), the estimates of gene expression that are used in this study are equal to the $\log_2$-transformed median of six probe signals per surveyed gene across six developmental stages.

Simulations

We tested the hypothesis that the degree of sequence similarity between the intra-IES 5′-end terminus and the reverse-complement of the 3′-end terminus is nonrandom. First, we estimated the site-specific nucleotide frequency downstream from the conserved 5′-TA-3′, at positions 3, 4, 5 (set 1) and 6, 7, 8 (set 2) within the 5′-end terminus of the surveyed group of IESs. Second, we generated two arrays of tri-nucleotide sequences, where each nucleotide at a given position is randomly sampled on the basis of its actual site-specific frequency, for each of the two sets. The generated arrays are the same size as the group of IESs under study. Third, we recorded a count for the number of times that the two arrays exhibit (i) triple mismatches, (ii) triple matches, (iii) double matches and (iv) single matches. The first and the second step were repeated 100 times and the average of the expected counts was compared to actual observations. We repeated the whole procedure to test whether the observed number of matches between the surveyed IESs' upstream and (the reverse complement of) downstream sequences (IES contiguous positions ±1, ±2, ±3 and ±4, ±5, ±6) is also nonrandom.

Logo analyses

We extracted twenty nucleotides upstream and downstream of each of the surveyed IESs. We used the RNA Structure Logo (26,27) to study the over- or under-representation of nucleotides within these sequences. RNA Structure Logo requires information on background nucleotide frequencies, which we estimated after concatenating the two sets of 20-nt sequences. In the logo, the height of each nucleotide represents its observed frequency relative to its expected frequency. Nucleotides whose frequency is less than expected are displayed upside down.

Statistics and comparative study of differently-sized IESs

Statistical analyses were carried out in the R environment (http://www.R-project.org/). Before we performed comparative analyses, e.g., intergenic versus intragenic IESs, we systematically controlled for possible size-differences between the surveyed IES categories. To accomplish this, we included IESs with a size ranging between the average sizes of the IES categories under study from our dataset. Following this step, IES sizes in the two subsets become statistically indistinguishable.

RESULTS

Excess of complementary base pairing between the termini of *P. tetraurelia* IESs

If IESs form loops, then an excess of complementary bases might be observed at the IES ends (Figure 1). The observed counts of triple and double matches between the intra-IES 5′-end terminus and the reverse-complement of the 3′-end terminus positions 3, 4, 5 and 6, 7, 8 are indeed higher than expected by chance (χ^2 test, uncorrected P-value < 0.001; Table 1). This is true when the whole set of IESs is studied (43 086 observations), and when IESs are separated into intergenic (8011 observations) and intragenic (35 075 observations) (Table 1). Moreover, this trend generally holds when we study exon-mapping (32 865 observations) or intron-mapping IESs (2210 observations) separately (Supplementary Table S1).

While significant overall, the deviations from random expectations are considerably more pronounced for intra-IES termini positions 3, 4, 5 compared to positions 6, 7, 8, in all cases (Table 1, Supplementary Table S1). For instance, the standardized residual relative to the number of triple matches in intragenic IESs is >11-fold higher for IES termini positions 3, 4, 5 compared to immediately downstream positions 6, 7, 8 (i.e. 42.1 versus 3.6). Furthermore, we observed that deviations from random expectations at IES termini positions 3, 4, 5 are more prominent for intragenic IESs compared to intergenic IESs. For instance, the ratio of the standardized residuals estimated for intragenic IES termini positions 3, 4, 5 and 6, 7, 8 is three-fold higher compared to the counterpart estimated for intergenic IESs (i.e. 11.7 versus 3.9). These observations indicate that complementary base pairing between the sub-terminal sequences of IESs is significantly, however not uniformly, elevated. Under the hypothesis that high levels of complementary base pairing facilitate the formation of DNA loops, our findings imply that (i) IES sub-terminal positions 3, 4, 5 play a more important role in loop formation compared to IES sub-terminal positions 6, 7, 8, and (ii) intragenic IESs might form loops more efficiently compared to intergenic IESs.

Excess of complementary base pairing between somatic DNA sequences abutting *P. tetraurelia* IESs

If IESs form loops, then an excess of complementary bases might also be observed at sites that are contiguous to IESs. We therefore studied the number of mismatches between IESs' immediately upstream and (the reverse complement of) downstream sequences. We found that the number of triple and double matches between the nucleotides that occupy IES-flanking sites ±1, ±2, and ±3 exceeds expectations significantly (χ^2 test, uncorrected P-value < 0.001; Table 1, Supplementary Table S1). At these sites, we also detected considerable levels of nucleotide over- and under-representation (Supplementary Figure S1A), which is in line with some of our previous findings (6). On the other hand, no or a marginal excess of triple or double matches were found when we analyzed the three more external sites ±4, ±5, ±6 (Table 1, Supplementary Table S1). These observations suggest that IESs may form loops by leveraging the excess of base complementarity between IES contiguous positions ±1, ±2 and ±3.

Significant positive relationship between C_{in}-score and C_{out}-score and IES length

It is known that IES length affects its excision. If the C_{in}-score and the C_{out}-score (Figure 1) truly plays a role in the process of DNA splicing, then one may expect to detect an association between these scores and IES length. Indeed, both the C_{in}-score and the C_{out}-score are significantly and positively correlated with IES length, though to different extents (Figure 2).

Table 1. Observed and expected number of mismatches between: (i) 5′-end and the reverse complement of 3′-end IES sub-terminal bases that occupy positions 3, 4, 5 or 6, 7, 8 downstream from the conserved 5′-TA-3′, and (ii) upstream and the reverse complement of downstream IES-flanking bases at positions 1, 2, 3 and 4, 5, 6

		Number of mismatches					
		0	1	2	3	χ^2	*P-value*
Intra-IES positions							
All IESs (pos. 3, 4, 5)	Exp.	4780	15726	16771	5809	4897.7	$<2.2 \times 10^{-16}$
	Obs.	9850	19922	10899	2415		
All IESs (pos. 6, 7, 8)	Exp.	2363	11982	19232	9509	108.4	$<2.2 \times 10^{-16}$
	Obs.	2754	12688	19158	8486		
Intergenic IESs (pos. 3, 4, 5)	Exp.	931	2965	3079	1037	758.8	$<2.2 \times 10^{-16}$
	Obs.	1847	3602	2077	485		
Intergenic IESs (pos. 6, 7, 8)	Exp.	390	2091	3596	1934	51.3	4.2×10^{-11}
	Obs.	536	2268	3546	1661		
Intragenic IESs (pos. 3, 4, 5)	Exp.	3832	12759	13717	4767	4171.0	$<2.2 \times 10^{-16}$
	Obs.	8003	16320	8822	1930		
Intragenic IESs (pos. 6, 7, 8)	Exp.	1991	9889	15631	7565	64.2	7.5×10^{-14}
	Obs.	2218	10420	15612	6825		
IES-flanking positions							
All IESs (pos. 1, 2, 3)	Exp.	1220	8280	18664	13930	590.9	$<2.2 \times 10^{-16}$
	Obs.	1954	10175	18429	11536		
All IESs (pos. 4, 5, 6)	Exp.	1024	7535	18473	15061	32.2	4.8×10^{-7}
	Obs.	1225	7888	18326	14655		
Intergenic IESs (pos. 1, 2, 3)	Exp.	303	1817	3565	2325	44.0	1.5×10^{-9}
	Obs.	433	2005	3460	2112		
Intergenic IESs (pos. 4, 5, 6)	Exp.	296	1788	3556	2370	13.5	0.004
	Obs.	390	1773	3511	2336		
Intragenic IESs (pos. 1, 2, 3)	Exp.	960	6592	15082	11450	492.6	$<2.2 \times 10^{-16}$
	Obs.	1521	8170	14969	9424		
Intragenic IESs (pos. 4, 5, 6)	Exp.	780	5893	14891	12521	7.8	0.05
	Obs.	835	6115	14815	12319		

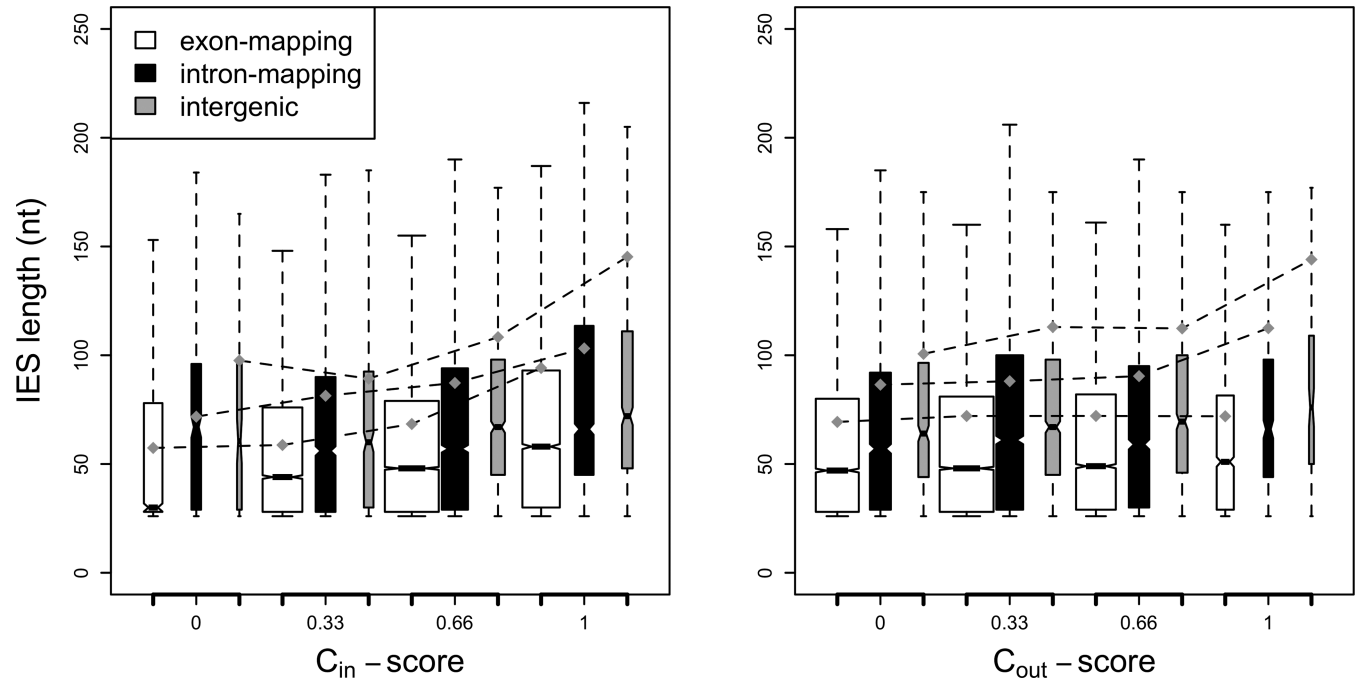

Figure 2. Boxplots illustrating the relationship between C_{in}-scores or C_{out}-scores and length (in nucleotides (nt)) of exon-mapping, intron-mapping, and intergenic IESs. Diamonds indicate average IES lengths. Boxes are drawn with widths proportional to the square roots of the number of observations in the groups.

The correlation between C_{out}-score and IES length is comparably weak for intragenic and intergenic IESs (Pearson's $r = 0.028$ versus 0.010, respectively; Fisher's Z; P-value $= 0.07$), whereas the overall stronger relationship between C_{in}-score and IES length is more robust for intragenic IESs compared to intergenic IESs (Pearson's $r = 0.092$ versus 0.065; Fisher's Z; P-value $= 0.014$). Also, we find that equivalent changes in the size of intergenic and intragenic IESs are associated with relatively greater increments in the C_{in}-score of intragenic IESs (ANOVA, $F = 67.7$, df $= 1$, P-value < 0.001).

With regard to IESs that reside within exons or introns, C_{in}-score (though not C_{out}-score) and IES size are also significantly and positively correlated both in the case of exon-mapping IESs (32 865 observations; Pearson's $r = 0.092$; P-value < 0.001) and for intron-mapping IESs (2210 observations; Pearson's $r = 0.073$; P-value < 0.001) (Figure 2). However, neither the strength of these linear relationships nor the slope of the corresponding regression lines differ between exon- and intron-mapping IESs (Fisher's Z and ANOVA, P-values > 0.05), in contrast to the case of intergenic and intragenic IESs.

Taken together, these data suggest that (i) C_{in}-score and C_{out}-score may indeed play a role in IES recognition/excision and (ii) the formation of DNA loops for increasingly larger IESs generally requires increasingly higher C_{in}-scores and, to a lesser extent, C_{out}-scores.

Intragenic IESs have higher C_{in}-scores than intergenic IESs

Provided that high levels of complementary base pairing between intra-IES termini positions 3, 4, 5 and between IES-flanking positions ±1, ±2, ±3 facilitate the formation of looped IESs, we asked how variations in the efficiency of DNA looping formation affect IES recognition/excision. To address this question we studied the C_{in}-scores and C_{out}-scores of intragenic and intergenic IESs. Once the different size distributions of these two IES populations are taken into consideration—intergenic IESs are significantly larger, on average [median], compared to intragenic IESs (111 [59] nt versus 72 [49] nt; Wilcoxon rank sum test, P-value < 0.001)—we expected to detect higher C_{in}-scores and C_{out}-scores for IESs that reside within, rather than outside, genes. The reason is that mutations such as sequence insertions in intergenic DNA regions (owing to imperfect IES excision) should be less likely to measurably affect fitness compared to mutations within genic sequences.

Indeed, when we examined the C-scores of intergenic and intragenic IESs whose size ranges between the averages of the two IES populations (72 nt and 111 nt), the 7334 intragenic IESs that were included in our subset show a significantly higher C_{in}-score compared to 1824 intergenic IESs (0.620 versus 0.588, respectively; Wilcoxon rank sum test, P-value < 0.001). Instead, we detect no significant difference for the C_{out}-scores of these two IES populations (0.360 versus 0.367, respectively; Wilcoxon rank sum test, P-value $= 0.451$). These results are in accordance with the notion that relatively higher levels of complementary base pairing between the IESs' sub-terminal sequences facilitate IES recognition/excision.

Intron- and Exon-mapping IESs have comparable C_{in}-scores and C_{out}-scores

We followed the same line of reasoning to ask how variations in the efficiency of DNA looping formation are associated with the recognition/excision of IESs mapping to the exons and to the introns of *P. tetraurelia*. Spliceosomal introns are generally assumed to evolve under no particular selective constraints. Moreover, because these intragenic noncoding sequences are typically excised from precursor mRNAs during the process of transcription, it is clear that accidental IES retention within excised introns would not affect protein primary structure. That noted, *P. tetraurelia* introns are exceptionally small (25 nt, on average (2,28)), which suggests that intron expansion owing to IES retention may be selectively disfavored.

We found that intron-mapping IESs display C-scores that are higher, on average, compared to those that we estimated for exon-mapping IESs (C_{in}-score: 0.651 versus 0.620; C_{out}-score: 0.375 versus 0.349; Wilcoxon rank sum test, P-value < 0.001). However, intron-mapping IESs are also larger, on average [median], compared to exon-mapping IESs (90 [67] nt versus 71 [48] nt; Wilcoxon rank sum test, P-value < 0.001). When we control for IES size by estimating the C_{in}- and C_{out}-scores of intron-mapping and exon-mapping IESs with overlapping lengths (see Materials and Methods), we found that the 2140 exon-mapping IESs that were included in our subset show C_{in}-scores that are comparable to those of 190 intron-mapping IESs (0.618 versus 0.619, respectively; Wilcoxon rank sum test, P-value $= 0.882$). Also, the highly significant difference that we had originally observed for the C_{out}-scores of these two populations is now only marginally significant (0.351 versus 0.389, respectively; Wilcoxon rank sum test, P-value $= 0.045$). In sum, the C-scores of intron-mapping IESs are comparable to the C-scores of exon-mapping IESs. It follows that in contrast to what is often theorized for other eukaryotic species, introns in *P. tetraurelia* might not tolerate sequence insertions better than exons.

Non-random variations in the relationship between C_{in}-score and C_{out}-score and IES length

Upon closer inspection, our dataset reveals a non-linear relationship between C-scores and IES length (Figure 3). When we focus on IESs that are up to 200nt large (95% of the surveyed elements), the C_{in}-score and the IES length jointly increase until the IES length is up to ∼62 nt (25 115 observations; Pearson's $r = 0.19$, P-value < 0.001), or larger than ∼108 nt (6235 observations; Pearson's $r = 0.08$, P-value < 0.001). In the intervening 62–108 nt interval, however, the C_{in}-score and the IES length are negatively correlated (11 736 observations; Pearson's $r = -0.07$, P-value < 0.001). On the other hand, the positive relationship between C_{out}-score and IES size essentially holds only for IESs that are up to ∼62 nt large (Pearson's $r = 0.05$, P-value < 0.001) (Figure 3).

We also detected non-negligible variations in the average C-scores of IESs that are smaller than 62 nt. In particular, IESs that are 26–30 nt or 44–45 nt large (17 346 observations) show a C_{in}-score that is significantly lower compared

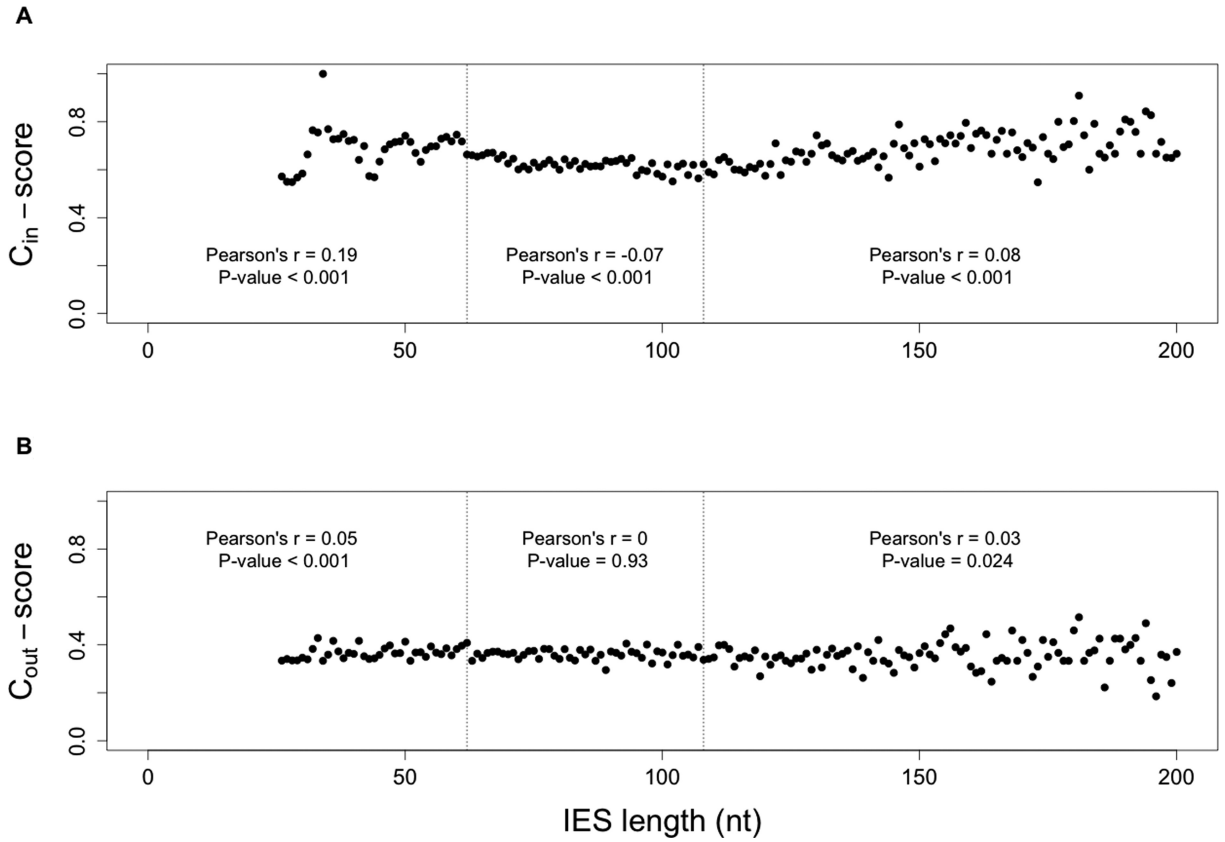

Figure 3. Scatterplots illustrating the relationship between IES length (in nucleotides (nt)) and C_{in}-score (**A**) or C_{out}-score (**B**) for IESs that are up to 200nt large (95% of the surveyed IESs). The overall correlation between C_{in}-score and IES length is positive and significant (43 086 observations; Pearson's $r = 0.077$, P-value < 0.001). The overall correlation between C_{out}-score and IES length is positive and significant (43 086 observations; Pearson's $r = 0.017$, P-value < 0.001). Horizontal dashed lines at IES length 62nt and 108nt demarcate subsections of the dataset for which the relationship between C_{in}-score and IES length shows the most apparent changes.

to IESs with an intervening size (i.e. 31–43 nt; 1073 observations) (0.563 versus 0.702, Wilcoxon rank sum test, P-value < 0.001 (Figure 4). These two IES populations show the opposite trend in terms of C_{out}-score (0.356 versus 0.338, respectively; Wilcoxon rank sum test, P-value $= 0.05$). It is worth noting that the IESs that are 31–43 nt large (hereinafter referred to as 'averted IESs') occur at a considerably low frequency in the *P. tetraurelia* germline genome (3). Moreover, the levels of complementary base pairing between intra-IES termini in this population only slightly exceed random expectations (χ^2 test, uncorrected P-values $=$ 0.04) (Supplementary Table S1).

Finally, we found that the average C_{in}-score increases abruptly when the IES length exceeds 45 nt, achieving values that are comparable to those of the averted IESs (averages; $C_{in}\text{-score}_{45nt} = 0.569$, $C_{in}\text{-score}_{46nt} = 0.634$, $C_{in}\text{-score}_{47nt} = 0.685$). A steady and significant increase was also detected for the C_{out}-score (averages; $C_{out}\text{-score}_{45nt} = 0.344$, $C_{out}\text{-score}_{46nt} = 0.358$, $C_{out}\text{-score}_{47nt} = 0.385$) (Figure 4).

The interplay between IES length, C_{in}-score and gene expression indicates that large IESs may be in a loop configuration prior to excision

It has been proposed that IESs larger than ∼45 nt are excised via a mechanism that involves the crosstalk between IES ends. On the other hand, smaller IESs—which are considered too short to form DNA loops—may rely on a distinct, presumably DNA loop-free mechanism of excision (3). Based on this proposal and on the observations described above, we hypothesized that the C-scores of IESs that are larger than 45 nt may be particularly elevated in a selective environment that prevents or minimizes the accumulation of harmful loop-perturbing variants. On the other hand, the C-scores of smaller IESs should stay low regardless of the selective environment to which the IESs are exposed as their excision might limitedly, if at all, rely on the formation of DNA loops (Supplementary Tables S1 and S2).

We tested this hypothesis by focusing on exon-mapping IESs that reside in protein-coding genes that are highly and weakly expressed in *P. tetraurelia*. Highly expressed genes evolve under stronger levels of selective pressure compared to weakly expressed genes (29). Therefore, large IESs residing in highly expressed genes should exhibit relatively higher C-scores compared to large IESs in weakly expressed

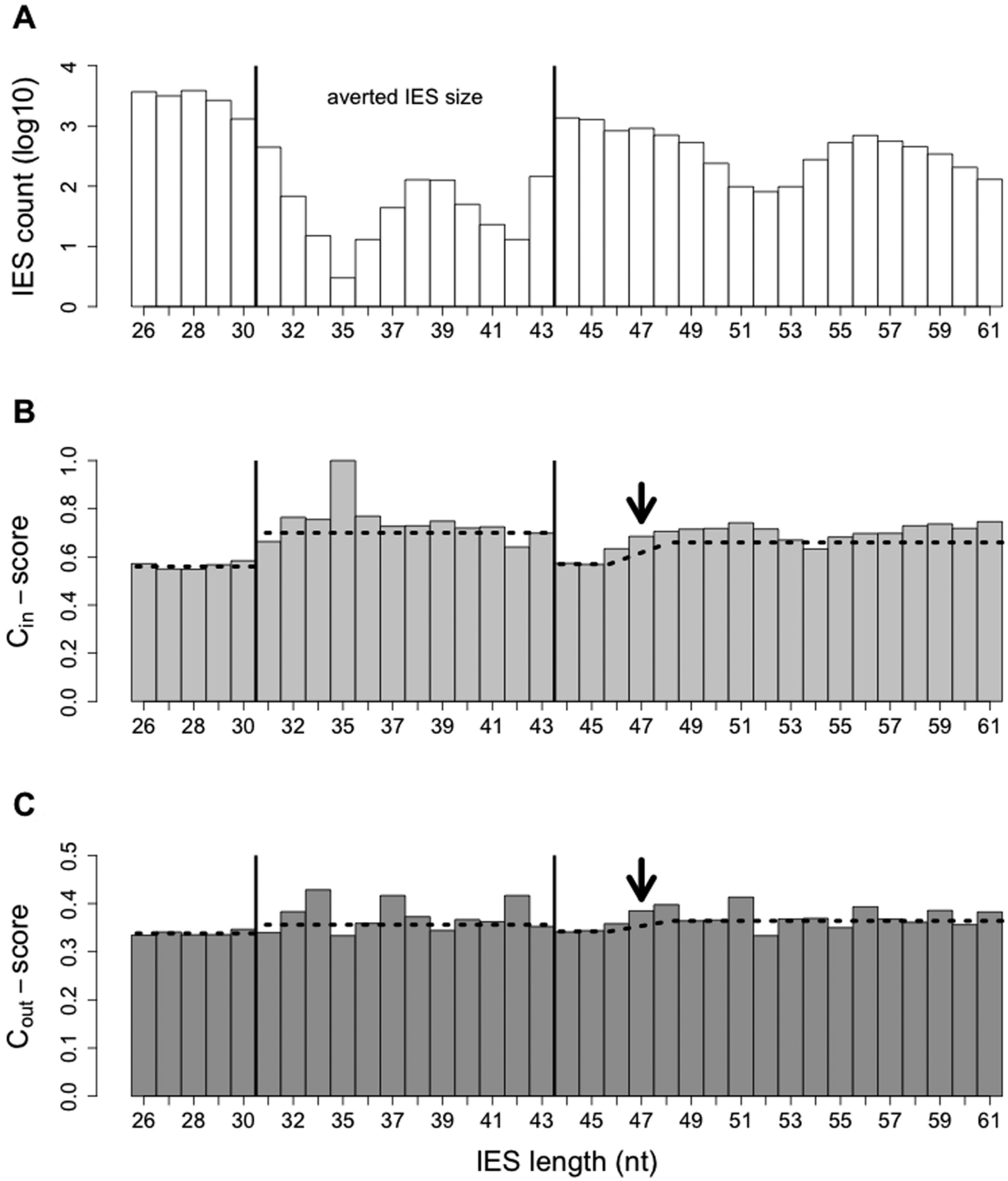

Figure 4. Barplots displaying the $\log_{10}$-transformed count (**A**), the average C_{in}-score (**B**), and the average C_{out}-score (**C**) of IESs ranging between 26nt and 61nt. Black horizontal dashed lines denote the average C_{in}-score or C_{out}-score calculated for groups of IESs with distinct lengths (i.e. 26–30nt; 31–43nt; 44–45nt; 46–61nt). Arrows denote the IES size-classes where a rapid and significant increase in average C_{in}- and C_{out}-score is detected. Note that the y-axis of the barplot in (**C**) has a max value of 0.5 and not 1 as in (**B**).

genes, all else being equal. In accordance with these expectations, the average C_{in}-score of large IESs mapping to highly expressed genes is significantly higher compared to the C_{in}-score of same size-class IESs mapping to weakly expressed genes (0.672 versus 0.651, respectively; Wilcoxon rank sum test, P-value < 0.001) (Figure 5A). This difference is even greater (i.e. 0.640 versus 0.584) when the distinct size of the IESs that reside in the two expression environments is factored in (see Materials and Methods) (IESs in lowly expressed genes: 123 nt; IESs in highly expressed genes: 91 nt; Wilcoxon rank sum test, P-value < 0.001). In contrast, small IESs with a size of 26–30 nt or 44–45 nt show comparable C_{in}-scores regardless of the expression levels of their gene of residence (0.559 versus 0.553, respectively; Wilcoxon rank sum test, P-value $= 0.442$) (Figure 5A). These findings lend support to a DNA loop-mediated excision of large IESs, but not of small IESs. With regard to averted IESs, the extent to which the C_{in}-score affects these IESs' excision is ambiguous. Within this size class, the IESs that reside in lowly expressed genes have a C_{in}-score that is marginally higher compared to averted IESs that occupy highly expressed genes (0.728 *vs* 0.683, respectively; Wilcoxon rank sum test, P-value < 0.05). After accounting for size, the observed difference might disappear however (the test cannot be performed due to the insufficient number of observations).

Finally, an elevated C_{out}-score was systematically detected for IESs that reside in highly expressed genes, irrespective of these IESs' size-class (Wilcoxon rank sum test, P-value < 0.001) (Figure 5B). These differences in C_{out}-score disappear when we restrict our analysis to IESs whose size ranges between the average sizes of the IESs residing in highly and weakly expressed genes (Wilcoxon rank sum test, P-value > 0.05). This finding suggests that unlike the C_{in}-score, increased C_{out}-scores might facilitate IES excision in an IES length-independent manner.

Small IESs in highly expressed genes may be recognized/excised less efficiently than large IESs

We observed that small IESs occur preferentially in weakly expressed genes, whereas large IESs (and averted IESs) reside most often in highly expressed genes ($\chi^2 = 241.09$, df $= 2$, P-value < 0.001) (Figure 5A). Because different gene expression environments impose distinct levels of selective pressure (29), the non-random distribution of IES-size classes hints that the mechanisms guiding the excision of large IESs and averted IESs is relatively more reliable compared to those guiding the excision of small IESs.

Although experimental evidence is required to support firmly this scenario, we reasoned that studying the relative number of IESs that theoretically activate the cellular surveillance systems when imperfectly excised might shed some light on this issue. Specifically, if imperfect IES excision—a potentially harmful event, particularly if in coding regions—truly occurs more often for small IESs than it does for other IESs, then small exon-mapping IESs should be more likely to contain premature termination codons (PTCs), as these elicit the Nonsense-Mediated Decay (NMD)-mediated removal of IES-retaining mRNAs.

A significant enrichment of PTC-containing small exon-mapping IESs was indeed detected in highly expressed genes relative to weakly expressed genes ($\chi^2 = 10.10$, df $= 1$, P-value < 0.01). In contrast, no overrepresentation was detected for PTC-containing large exon-mapping IESs ($\chi^2 = 2.1$, df $= 1$, P-value $= 0.148$). When we examined intronic IESs, we also found an even distribution of PTCs that was independent of IES size or gene expression level (χ^2 test, P-value > 0.05). Furthermore, we detected a relative excess of PTC-containing small (but not large) $3n$ IESs in highly expressed genes ($\chi^2 = 7.4$, df $= 1$, P-value < 0.01). This finding is in line with the hypothesis that small IESs are imperfectly excised more often than are large IESs, because transcripts retaining PTC-free IESs with a size that is multiple of 3 (i.e. $3n$) are invisible to NMD. Finally, we found that PTC-containing IESs, both small and large, reside at the 5′ end of highly expressed genes more often than at these genes' 3′ end (small IESs: $\chi^2 = 5.4$, df $= 1$, P-value $= 0.02$; large IESs: $\chi^2 = 9.8$, df $= 1$, P-value < 0.01). No positional bias was detected however for PTC-containing IESs occupying weakly expressed genes (small IESs: $\chi^2 = 2.8$, df $= 1$, P-value $= 0.09$; large IESs: $\chi^2 = 0.006$, df $= 1$, P-value $= 0.94$). These observations are compatible with *in silico* evidence indicating that NMD in *P. tetraurelia* removes PTCs that reside toward the gene 5′ end most efficiently (6). That noted, we further found that PTC-free IESs reside also preferentially at the gene 5′ end compared to the 3′ end (small IESs: $\chi^2 = 22.5$, df $= 1$, P-value < 0.001; large IESs: $\chi^2 = 22.9$, df $= 1$, P-value < 0.001). This overall positional bias toward the gene 5′ end suggests that IESs in *P. tetraurelia* may be most often gained at this location and/or preferentially lost from the gene 3′ end.

DISCUSSION

This study provides several original insights into the molecular mechanisms and the evolutionary processes that are associated with programmed DNA deletion in the ciliate Paramecium.

In addressing the first question we posed—'What is the likelihood that *P. tetraurelia* IESs may form loops?'—we found that the levels of base complementarity between IES termini and between IES-flanking positions deviate significantly from expectations of randomly distributed matches (Table 1, Supplementary Tables S1 and S2). This finding lends support to the hypothesis that some fraction of IESs may be in a loop configuration prior to excision. The levels of base complementarity between IES termini positions 3, 4, 5 (C_{in}-score) and between IES-flanking positions ±1, ±2 and ±3 (C_{out}-score) are considerably higher compared to the levels of base complementarity that we estimated for IES termini positions 6, 7, 8 and IES-flanking positions ±4, ±5 and ±6, respectively (Table 1, Supplementary Table S1). We interpret these findings as demarcating the identity of *cis*-acting sequences, within and outside IESs, which may be most critical in assisting in and/or stabilizing the formation of a DNA loop. Consistent with the importance of these sites in IES recognition/excision, we recently reported that bases occupying IES termini positions 3, 4, 5 and IES-flanking positions ±1 and ±2 show considerable levels of conservation across three Paramecium species (*P.*

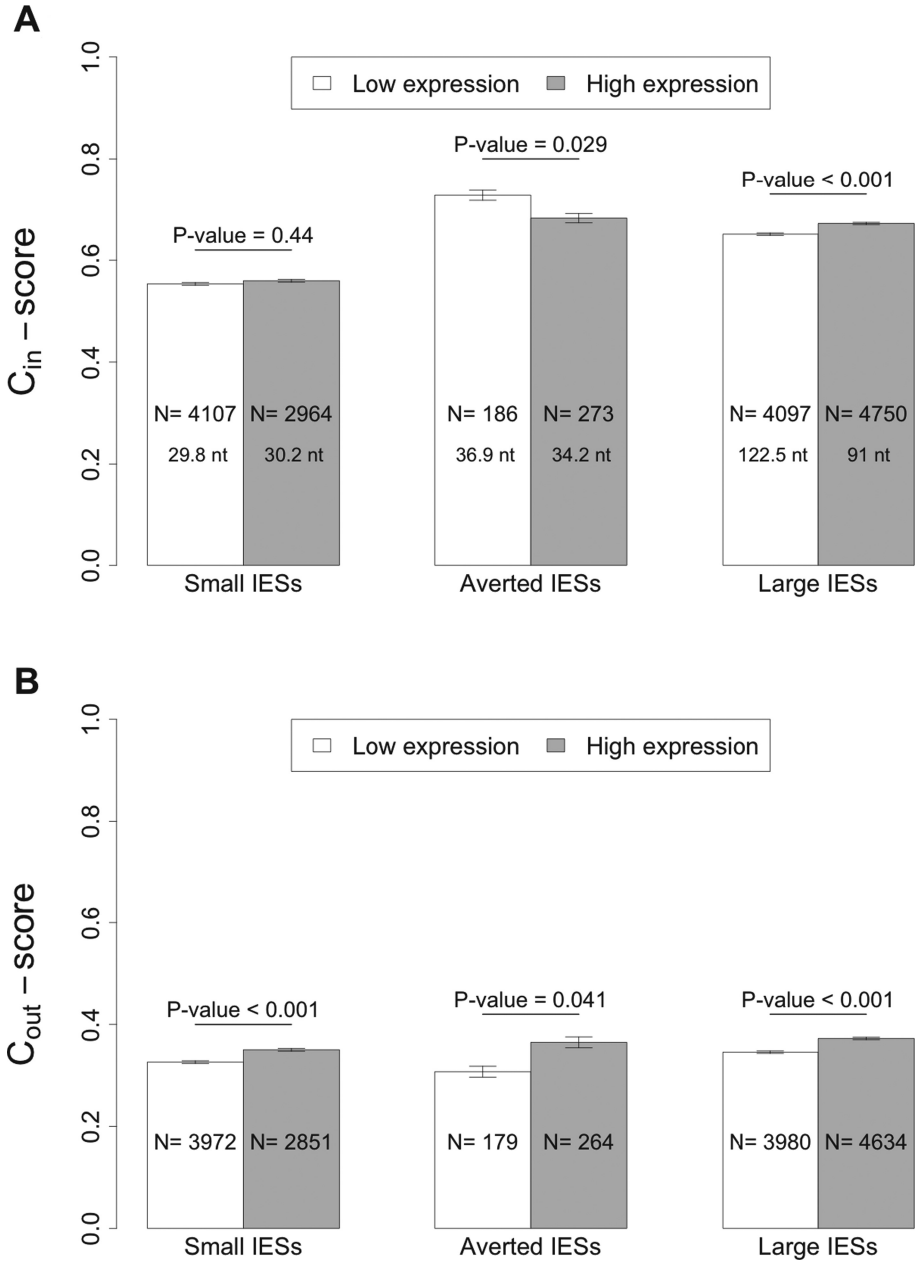

Figure 5. Barplots displaying the average values (with corresponding standard errors) of C_{in}-score (**A**) and C_{out}-score (**B**) of small IESs (26–30nt and 44–45nt), averted IESs (31–43nt), and large IESs (>45nt) residing in weakly and highly expressed genes. Low and high expression values occupy the lower and the upper quartile of a distribution of gene expression values (see Materials and Methods). For each group of IESs, Wilcoxon rank sum test *P*-values are denoted together with the number (N) and the average size (in nucleotides (nt)) of the surveyed IESs.

tetraurelia, *P. biaurelia*, and *P. sexaurelia*). Furthermore, the degree of sequence preservation at these positions is more pronounced in frequently excised sequences (i.e. imperfectly excised IESs) than it is in occasionally excised somatic DNA sequences (i.e. cryptic IESs) (6).

Our second question—'Do variations in complementary base pairing between IESs' sub-terminal or contiguous sequences affect the efficiency of IES recognition/excision?'—logically follows the first one. We found that the C_{in}-score and the C_{out}-score are associated with IES length (Figure 2), a property that is known to play a critical role in the process of IES excision (9,16,21). The strength and the sign of the observed relationships, however, may differ between genomic locations (Figure 2) and vary with IES length (Figure 3). In particular, while the relationship between the C_{in}-score and IES length is relatively robust—more so for intragenic IESs than for intergenic IESs—and easily detectable across all IES sizes, it is not consistently positive (Figure 3). In contrast, the association between C_{out}-score and IES length is weak for intragenic and intergenic IESs alike, while it is significant and positive for short IESs only (Figure 3). These differences suggest that C_{in}- and C_{out}-scores might affect the process of IES recognition/excision in different ways.

In addition, we found that the C_{in}-score and the C_{out}-score are generally increased in genomic regions where one expects that intense natural selection minimize the generation of imperfect, potentially harmful, IES excisions. For example, the C_{in}-score of intragenic IESs is higher compared to the C_{in}-score of intergenic IESs. Importantly, this difference in C_{in}-score only emerges after the differences in the size distribution between the two IESs populations have been accounted for.

Finally, C_{in}-score and C_{out}-score increase together, rapidly and significantly, as soon as IES length exceeds 45 nt (Figure 4). Together with the rest of our findings, this observation suggests a unifying framework for the mechanisms of IES excision in *P. tetraurelia*, which is in line with and extends previous models (8,19,21). Specifically, during assembly of the excision complex, the increased C-scores (C_{in}-score, in particular) of large IESs (>45-nt) assists in and/or stabilizes the formation of DNA loops, possibly under the form of 4-stranded or cruciform DNA structures, thereby facilitating cleavage. Alternatively, complementary sequences at each IES end might favor the assembly of a symmetric protein-DNA complex that catalyzes IES excision. Regardless of the exact mechanism through which complementary bases at large IESs' ends may operate, small RNAs and histone modifications probably promote such mechanism (9,16). In contrast, the excision of small IESs (26–30 nt and 44–45 nt), whose C_{in}-scores are both relatively weak (Figure 4, Supplementary Table S2) and unresponsive to distinct levels of selective pressures (Figure 5), may not involve the formation of DNA loops. It follows that small RNAs and histone modifications may not be as critical for promoting the excision of these small IESs as they are for the excision of large IESs, consistent with previous work (9,16). Under these circumstances, the PiggyMAC-mediated excision of small IESs would largely rely on the quality of *cis*-acting signals. Finally, the mechanism(s) guiding the recognition/excision of averted IESs (31–43 nt) remains uncertain at this point, owing to the ambiguous features of these sequences. For example, although these IESs are often flanked by nucleotides that are as recurrent in small IESs (Supplementary Figure S1B), they also exhibit C_{in}-scores and C_{out}-scores that closely resemble those of large IESs (Figure 4). In the face of these ambiguities, our current observations do not support the notion that these IESs are less efficiently excised compared to small or large IESs (3).

With regard to the third, and last question we posed—'Are the mechanisms guiding the excision of large IESs more reliable compared to those guiding the excision of small IESs?'—our results suggest that large IESs may be more efficiently excised compared to small IESs, at least in highly expressed genes. These observations do not entirely reconcile with the proposition that large IESs are preferentially removed by selection over time (21). In fact, this latter suggestion conflicts with one of our main observations, i.e. small (but not large) IESs are underrepresented in highly expressed coding sequences (Figure 5). Our interpretation for this deficit is that small IESs are less reliably recognized/excised compared to large IESs. Since it is potentially harmful, flagging imperfect excision to the cell would be advantageous, particularly when IESs occur within coding regions. Consistent with this rationale, small (but not large) IESs mapping to highly expressed exons are enriched with in-frame stops, which makes IES-retaining transcripts a favored target of the cellular surveillance systems.

If small IESs are truly less efficiently excised compared to large IESs, how could small IESs, whose imperfect excision is presumably disadvantageous, accumulate in the Paramecium genome and be preserved over evolutionary time? One possible explanation may be that losing small IESs is more difficult than losing large IESs. Although this is an as yet unsubstantiated speculation, it is worth noting that this condition would mimic an evolutionary stability *and* account for our findings. More explicitly, such condition could explain why small IESs are so copious in the Paramecium genome and evolutionary old (3). It could also explain why, rather than losing these IESs entirely, NMD-mediated defense mechanisms have evolved to minimize the potential harm resulting from their imperfect excision.

Finally, one important result of our investigation is that intron-mapping and exon-mapping IESs show C_{in}- and C_{out}-scores that are comparable, even after accounting for size. This finding implies that spliceosomal introns in *P. tetraurelia* evolve under some non-trivial selective constraints, which prevent the accumulation of sequence insertions. As hinted in the results section, this suggestion is quite reasonable given the exceptionally small size of spliceosomal introns in Paramecium. It is worth emphasizing that in examining IESs (the substrate of DNA splicing) we have, in this case, uncovered a property of spliceosomal introns (the substrate of RNA splicing). This connection is of particular interest as IESs and introns, two separate classes of noncoding sequences whose mechanisms of excision are entirely distinct, have already been reported to share a number of features (6, 30). Indeed, our study reveals three novel properties that these sequences have in common. First, for both DNA splicing and mRNA splicing, the quality of *cis*-acting excision signals scales positively with the size of the excised sequences (this study, (31–33)). Second, IESs and spliceosomal introns reside preferentially at the gene 5′ end (this study, (34)), a bias that may result from preferential loss from the gene 3′ end and/or preferential gain at the gene 5′ end (35). And third, the metric that we use to assess the quality of *cis*-acting IES recognition/excision signals is reminiscent of a measure commonly employed to estimate the quality of mRNA splicing signals, i.e. the degree of sequence complementarity between acceptor (or donor) sites and cognate small nuclear RNAs (36). Future studies might extend this list of similarities, and shed light on the significance, if any, of the analogy between IESs and spliceosomal introns.

CONCLUSION

Several published observations suggest that programmed DNA deletion in *P. tetraurelia* largely relies on the underlying genetic properties of excised germline sequences. Our study lends further support to these findings, extends the number of *cis*-acting regulatory elements, and reveals properties that most likely affect or are associated with the efficiency of the DNA deletion process. Not least, our

study supplies a novel quality measure for *cis*-acting IES recognition/excision signals, a metric whose further application should have far-reaching implications for the study of programmed DNA elimination in *Paramecium*.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank K. Karczewski, H.D. Görtz, and two anonymous reviewers for their valuable comments. Support by the Münster Graduate School of Evolution (MGSE) to D.F. and G.L. is gratefully acknowledged.

FUNDING

University of Münster and the Deutsche Forschungsgemeinschaft [CA1416/1-1 to F.C.]. Funding for open access charge: Deutsche Forschungsgemeinschaft [CA1416/1-1 to F.C.].
Conflict of interest statement. None declared.

REFERENCES

1. Sonneborn,T.M. (1975) The *Paramecium-Aurelia* complex of 14 sibling species. *Trans. Am. Microsc. Soc.*, **94**, 155–178.
2. Aury,J.M., Jaillon,O., Duret,L., Noel,B., Jubin,C., Porcel,B.M., Segurens,B., Daubin,V., Anthouard,V., Aiach,N. *et al.* (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, **444**, 171–178.
3. Arnaiz,O., Mathy,N., Baudry,C., Malinsky,S., Aury,J.M., Denby Wilkes,C., Garnier,O., Labadie,K., Lauderdale,B.E., Le Mouel,A. *et al.* (2012) The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet.*, **8**, e1002984.
4. Betermier,M., Duharcourt,S., Seitz,H. and Meyer,E. (2000) Timing of developmentally programmed excision and circularization of *Paramecium* internal eliminated sequences. *Mol. Cell. Biol.*, **20**, 1553–1561.
5. Duret,L., Cohen,J., Jubin,C., Dessen,P., Gout,J.F., Mousset,S., Aury,J.M., Jaillon,O., Noel,B., Arnaiz,O. *et al.* (2008) Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline. *Genome Res.*, **18**, 585–596.
6. Catania,F., McGrath,C.L., Doak,T.G. and Lynch,M. (2013) Spliced DNA sequences in the *Paramecium* germline: their properties and evolutionary potential. *Genome Biol. Evol.*, **5**, 1200–1211.
7. Baudry,C., Malinsky,S., Restituito,M., Kapusta,A., Rosa,S., Meyer,E. and Betermier,M. (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev.*, **23**, 2478–2483.
8. Gratias,A. and Betermier,M. (2003) Processing of double-strand breaks is involved in the precise excision of paramecium internal eliminated sequences. *Mol. Cell. Biol.*, **23**, 7152–7162.
9. Lhuillier-Akakpo,M., Frapporti,A., Denby Wilkes,C., Matelot,M., Vervoort,M., Sperling,L. and Duharcourt,S. (2014) Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting determinants for zygotic genome rearrangements. *PLoS Genet.*, **10**, e1004665.
10. Kapusta,A., Matsuda,A., Marmignon,A., Ku,M., Silve,A., Meyer,E., Forney,J.D., Malinsky,S. and Betermier,M. (2011) Highly precise and developmentally programmed genome assembly in Paramecium requires ligase IV-dependent end joining. *PLoS Genet.*, **7**, e1002049.
11. Marmignon,A., Bischerour,J., Silve,A., Fojcik,C., Dubois,E., Arnaiz,O., Kapusta,A., Malinsky,S. and Betermier,M. (2014) Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *PLoS Genet.*, **10**, e1004552.
12. Arambasic,M., Sandoval,P.Y., Hoehener,C., Singh,A., Swart,E.C. and Nowacki,M. (2014) Pdsg1 and Pdsg2, novel proteins involved in developmental genome remodelling in Paramecium. *PLoS One*, **9**, e112899.
13. Duharcourt,S., Butler,A. and Meyer,E. (1995) Epigenetic self-regulation of developmental excision of an internal eliminated sequence on *Paramecium tetraurelia*. *Genes Dev.*, **9**, 2065–2077.
14. Duharcourt,S., Keller,A.M. and Meyer,E. (1998) Homology-dependent maternal inhibition of developmental excision of internal eliminated sequences in *Paramecium tetraurelia*. *Mol. Cell. Biol.*, **18**, 7075–7085.
15. Lepere,G., Nowacki,M., Serrano,V., Gout,J.F., Guglielmi,G., Duharcourt,S. and Meyer,E. (2009) Silencing-associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. *Nucleic Acids Res.*, **37**, 903–915.
16. Sandoval,P.Y., Swart,E.C., Arambasic,M. and Nowacki,M. (2014) Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting. *Dev. Cell*, **28**, 174–188.
17. Klobutcher,L.A. and Herrick,G. (1995) Consensus inverted terminal repeat sequence of *Paramecium* IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons. *Nucleic Acids Res.*, **23**, 2006–2013.
18. Mayer,K.M. and Forney,J.D. (1999) A mutation in the flanking 5′-TA-3′ dinucleotide prevents excision of an internal eliminated sequence from the *Paramecium tetraurelia* genome. *Genetics*, **151**, 597–604.
19. Gratias,A., Lepere,G., Garnier,O., Rosa,S., Duharcourt,S., Malinsky,S., Meyer,E. and Betermier,M. (2008) Developmentally programmed DNA splicing in *Paramecium* reveals short-distance crosstalk between DNA cleavage sites. *Nucleic Acids Res.*, **36**, 3244–3251.
20. Mayer,K.M., Mikami,K. and Forney,J.D. (1998) A mutation in *Paramecium tetraurelia* reveals functional and structural features of developmentally excised DNA elements. *Genetics*, **148**, 139–149.
21. Swart,E.C., Wilkes,C.D., Sandoval,P.Y., Arambasic,M., Sperling,L. and Nowacki,M. (2014) Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Res.*, **42**, 8970–8983.
22. Arnaiz,O., Cain,S., Cohen,J. and Sperling,L. (2007) ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. *Nucleic Acids Res.*, **35**, D439–D444.
23. Arnaiz,O. and Sperling,L. (2011) ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*. *Nucleic Acids Res.*, **39**, D632–D636.
24. Arnaiz,O., Gout,J.F., Betermier,M., Bouhouche,K., Cohen,J., Duret,L., Kapusta,A., Meyer,E. and Sperling,L. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate *Paramecium tetraurelia*. *BMC Genomics*, **11**, 547.
25. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
26. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
27. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.*, **13**, 583–586.
28. Jaillon,O., Bouhouche,K., Gout,J.F., Aury,J.M., Noel,B., Saudemont,B., Nowacki,M., Serrano,V., Porcel,B.M., Segurens,B. *et al.* (2008) Translational control of intron splicing in eukaryotes. *Nature*, **451**, 359–362.
29. Gout,J.F., Kahn,D., Duret,L. and Paramecium Post-Genomics,C. (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.*, **6**, e1000944.
30. Catania,F. and Schmitz,J.(2015) On the path to geneticnovelties: insights from programmed DNA elimination and RNA splicing. *RNA*, doi:10.1002/wrna.1293.
31. Weir,M. and Rice,M. (2004) Ordered partitioning reveals extended splice-site consensus information. *Genome Res.*, **14**, 67–78.
32. Fahey,M.E. and Higgins,D.G. (2007) Gene expression, intron density, and splice site strength in Drosophila and Caenorhabditis. *J. Mol. Evol.*, **65**, 349–357.

33. Dewey,C.N., Rogozin,I.B. and Koonin,E.V. (2006) Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics*, **7**, 311.
34. Lin,K. and Zhang,D.Y. (2005) The excess of 5′ introns in eukaryotic genomes. *Nucleic Acids Res.*, **33**, 6522–6527.
35. Catania,F. and Lynch,M. (2008) Where Do Introns Come From? *PLoS Biol.*, **6**, e283.
36. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.