In silico analysis of the sequence features responsible for alternatively spliced introns in the model green alga Chlamydomonas reinhardtii

Praveen-Kumar Raj-Kumar, Olivier Vallon, Chun Liang

In silico analysis of the sequence features responsible for alternatively spliced introns in the model green alga *Chlamydomonas reinhardtii*

Praveen Kumar Raj Kumar[1,2§], Olivier Vallon[3], Chun Liang[1§]

[1]Department of Biology, Miami University, Oxford, Ohio 45056, USA
[2]Current address: Chan Soon-Shiong Institute of Molecular Medicine at Windber, Windber, Pennsylvania 15963, USA
[3]Institut de Biologie Physico-Chimique, UMR 7141 CNRS/Université Pierre et Marie Curie, 13 rue Pierre et Marie Curie, 75005 Paris, France

[§]Co-corresponding author

Email addresses:
PKRK: rajkump@miamioh.edu
OV: ovallon@ibpc.fr
CL: liangc@miamioh.edu

Abstract

Alternatively spliced introns are the ones that are usually spliced but can be occasionally retained in a transcript isoform. They are the most frequently used alternative splice form in plants (~50% of alternative splicing events). *Chlamydomonas reinhardtii*, a unicellular alga, is a good model to understand alternative splicing (AS) in plants from an evolutionary perspective as it diverged from land plants a billion years ago. Using over 7 million cDNA sequences from both pyrosequencing and Sanger sequencing, we found that a much higher percentage of genes (~20% of multi-exon genes) undergo AS than previously reported (3-5%). We found a full component of SR and SR-like proteins possibly involved in AS. The most prevalent type of AS event (40%) was retention of introns, most of which were supported by multiple cDNA evidence (72%) while only 20% of them have coding capacity. By comparing retained and constitutive introns, we identified sequence features potentially responsible for the retention of introns, in the framework of an "intron definition" model for splicing. We find that retained introns tend to have a weaker 5' splice site, more Gs in their poly-pyrimidine tract and a lesser conservation of nucleotide ‘C’ at position -3 of the 3' splice site. In addition, the sequence motifs found in the potential branch-point region differed between retained and constitutive introns. Furthermore, the enrichment of G-triplets and C-triplets among the first and last 50 nt of the introns significantly differ between constitutive and retained introns. These could serve as intronic splicing enhancers. All the alternative splice forms can be accessed at http://bioinfolab.miamioh.edu/cgi-bin/PASA_r20140417/cgi-bin/status_report.cgi?db=Chre_AS

Introduction

RNA splicing is a process in which organisms precisely remove introns and join exons on a primary transcript. It is most prevalent in eukaryotes, where most mRNAs undergo splicing. Even though splicing happens with high fidelity, occasional retention of regions normally spliced as introns has been reported in many organisms. Retention of introns is only one of five major types of alternative splicing (AS), the others being exon skipping, alternative exon and alternative choices of 5' or 3' splice sites (ss) (Matlin et al. 2005; Reddy 2007). Retention of introns has been reported to alter the biological activity of many transcript isoforms (Carvalho et al. 2013; Ge and Porse 2014). For example, an intron retained in the *Arabidopsis* INDERMINATE DOMAIN 14 (IDD14) transcription factor was shown to enhance cold adaptation (Seo et al. 2011), and an intron retained in the transcripts involved in human haematopoietic stem cell was found to regulate granulocyte differentiation (Wong et al. 2013). Several retained intron events are also reported as targets of the non-sense mediated RNA decay (NMD) pathway (Carvalho et al. 2013; Ge and Porse 2014). For instance, introns retained in the *Arabidopsis FCA-α* transcript (Macknight et al. 2002; Quesada et al. 2003) and Human *Robo3.2* transcript (Colak et al. 2013) introduce a premature termination codon that targets them for NMD.

The percentage of AS events involving intron retention is vastly different in plants and animals. In plants, the genome-wide studies for Arabidopsis (Marquez et al. 2012), rice (Zhang et al. 2010) and cucumber (Guo et al. 2010) have revealed 40%, 47% and 54.4% respectively, of the AS events as intron retention. In contrast, in animals, the genome-wide studies for human (Pan et al. 2008) and mouse (Khodor et al. 2012) have found that less than 5% of the AS events as retention of introns. This disparity can be explained by the difference in the exon-intron architecture. Animals, especially humans, have short exons (average size ~ 170 nt) interspersed with long introns (average size ~ 5,000 nt) (Zhang 1998; Sakharkar et al. 2005). In contrast,

plants like *Arabidopsis* and rice possess much shorter introns (average sizes ~ 173 – 433 nt), while exons remain short (average sizes 172-193 nt) (Wang and Brendel 2006; Reddy 2007).

Two models of splice site recognition have been proposed to explain how short and long introns are spliced. Berget (Talerico and Berget 1994; Berget 1995) proposed an intron-defined splicing mechanism ("intron definition"), whereby serine/arginine-rich (SR) proteins of the spliceosome directly recognize a intronic splicing enhancer (ISE), and the accompanying splicing machinery spans across the intron to excise it (Reddy 2007; De Conti et al. 2013). This mechanism is believed to apply to introns that are short relative to flanking exons. Through *in vitro* splicing assays on Drosophila, it has been found that introns shorter than 250 nt are target for "intron definition"(Fox-Walsh et al. 2005). In short introns, the presence of either a weak ISE or an intronic splicing suppressor (ISS) can result in intron retention (Reddy 2007; De Conti et al. 2013). The higher frequency of intron retention in invertebrates has also been attributed to the presence of shorter introns (67-159 nt) (Sammeth et al. 2008). Talerico and Berget (Talerico and Berget 1994) observed that the short introns of *Drosophila* are spliced by an intron-defined mechanism and argued that intron retention can be expected if intronic splice sites are not properly recognized.

For short exons with long flanking introns (> 250 nt), an exon-defined splice mechanism has been proposed, where exonic splicing enhancer (ESE) motifs in exons invite SR proteins followed by the other proteins of the splicing machinery, to excise the flanking introns (Keren et al. 2010; De Conti et al. 2013). Here again, the presence of a weak ESE or of an exonic splicing suppressor (ESS) is believed to lead to exon skipping.

Genome-wide studies linking intron retention to the presence of weak intronic signals have been conducted for only a few model species such as human, *Drosophila* and *Arabidopsis* (Lim and Burge 2001; Sakabe and Souza 2007; Mao et al. 2014). In order to understand the

evolutionary patterns of AS and in particular the origin of intron retention in plants, it is necessary to characterize retained introns in more distantly-related taxa. In this study, we focused on the green alga, *Chlamydomonas reinhardtii*, an intron-rich, unicellular model organism for which both a genome sequence and large transcriptomics datasets are available. *C. reinhardtii* is a highly evolved representative of the green algae (Chlorophyta), the sister group of land plants (Streptophyta) in the Viridiplantae (green photosynthetic organisms) lineage. Based on 252,484 Sanger-ESTs, Labadorf et al. (Labadorf et al. 2010) conducted a genome-wide AS analysis in *C. reinhardtii* and showed that the extent of AS is much lower than that in land plants (3%). Incorporating a much larger dataset (7,345,432 cDNA sequences) generated by both Sanger and pyrosequencing technologies, our study was aimed at re-evaluating the prevalence of AS and exploring the sequence features responsible for intron retention in *C. reinhardtii.* Using GMAP (Wu and Watanabe 2005), BLAT (Kent 2002) and the Program to Assemble Spliced Alignment (PASA) (Haas et al. 2003; Campbell et al. 2006) to identify splice variants, and incorporating existing Phytozome (v9.0, http://www.phytozome.net/) gene models for this alga (Chre v5.3.1), we found that a much high percentage of multi-exon genes (3,342 genes, i.e. 19.9%) undergo AS. As in other plants analyzed so far, we found that the dominant mode of AS is intron retention (39.8% of AS genes). We also observed that in comparison with constitutively spliced introns (i.e., those that are always spliced in all the isoforms), retained introns (i.e., the ones that are retained in some isoforms) have comparatively weaker splice sites, fewer ISEs such as GGG and a different nucleotide composition in the putative Branch point region.

Results and Discussion

Detection of AS forms using PASA and integration with Phytozome gene models

We combined publicly available Sanger (338,243) and pyrosequencing (7,696,737) reads (See Methods) for *C. reinhardtii* to study the full extent of AS. To effectively use the PASA software (Haas et al. 2003; Campbell et al. 2006) for detection of AS, we first assembled the cDNA sequences with the guidance of genome mapping using CLC (http://www.clcbio.com/), yielding 62,234 contig sequences. The contigs were re-mapped to the *C. reinhardtii* genome (Chre v5.3.1) using two programs, GMAP (Wu and Watanabe 2005) and BLAT (Kent 2002) to capture all the possible high-quality alignments. The resulting spliced alignments were assembled and labeled for alternative splicing using the PASA software.

We used the mRNAs in the Chre v5.3.1 annotation, which already contains alternative transcripts for 1,789 loci (~ 10%), as the full length cDNAs in PASA to guide the alignment towards the gene loci. Here PASA produced 19,844 subclusters called the PASA genes. Because PASA considers only contig-to-genome alignment evidence and ignores other important criteria like homology and proteomics data that can validate a structure, incomplete cDNA coverage may lead PASA to split a single gene into two or more assemblies. Thus, the general tendency in PASA is to overestimate the number of predicted genes (19,844 PASA loci vs. 17,737 loci in Chre v5.3.1). To take full advantage of the high quality Augustus-based annotation of the *C. reinhardtii* genes (Blaby et al. 2014), we therefore utilized the *Annotation Comparison and Updates* option of PASA to update the PASA alignment assemblies by comparison to preexisting *C. reinhardtii* gene annotations whenever possible. Incorporation of existing Chre v5.3.1 annotation involves extending 3'/5' UTRs, expanding ORFs and consolidating novel alternatively spliced isoforms for a given locus based on PASA alternative splice isoform predictions. 17,701

loci in the Chre v5.3.1 annotation (out of 17,737) could thus be incorporated into the PASA models. In total, after consolidation, 13,680 mRNAs of Chre v5.3.1 gene models have been updated, while 647 new mRNAs were uncovered by PASA (Additional File 1 contains a gene model in 'gff3' format consolidating Chre v5.3.1 and the PASA updates). Among these gene models, 3,342 were identified by PASA as showing AS, which includes 370 genes described as alternatively spliced by the Augustus annotation but for which our EST dataset did not provide evidence for AS (Supplemental Figure 1). This represents ~20.0 % of the 16,743 multi-exon genes (http://bioinfolab.miamioh.edu/cgi-bin/PASA_r20140417/cgi-bin/alt_splice_report.cgi?db=Chre_AS). Figure 1a shows a snapshot of a PASA assembly showing intron retention together with exon skipping, as viewed in our PASA web portal. Among AS genes, 1,331 genes (39.8%) used intron retention, the largest of all AS categories (Fig.1a). Among the 5,198 AS events, 1,521 (29%) were retained introns (Fig.1b). Changes in acceptor site position (delta) were usually small, with peaks at 3, 6 and 9 (Fig. 2a), suggesting selection for frame conservation. A similar period-3 oscillation was also visible for changes in donor site position. At position 4, a large signal obscured this periodicity. We think that it is due to the prevalence of GT at positions 5-6 of the 5'ss consensus, the same residues as at positions 1-2 (Fig. 3a), which could favor an erroneous choice of the +4 position for splicing.

Our analysis captured most of the AS cases described in the literature, for example *CGE1* (Schroda et al. 2001) in PASA gene/subcluster 16859, ANK22 (Li et al. 2003) in subcluster 12201 and CTH1 (Moseley et al. 2002) in subcluster 3699. The AS event that has been described in the serine/arginine-rich (SR) protein homolog (Kalyna et al. 2006) (Cre02.g099350; see Supplemental Table 1) was also detected. We failed to detect AS of CCM1 (Fukuzawa et al. 2001) and the long 5' extension described in FLU1 (Falciatore et al. 2005), most likely due to lack of cDNA coverage. Finally, the hypothetical AS gene described as "Cr002" in (Li et al.

2003) was found to correspond to a highly repeated region in the genome, and was therefore not in our list.

Overall, our data shows a widespread occurrence of AS in *C. reinhardtii*, higher than previously reported by Labadorf et al. (Labadorf et al. 2010). An over twenty-fold increase in sequencing depth and coverage by pyrosequencing enabled us to identify 2,844 more genes that undergo AS (3,342 vs. 498) and 5,229 more alternative splice isoforms (5,840 vs. 611). Similarly, in *Arabidopsis thaliana*, a genome of comparable size (~135 Mb vs. ~120 Mb for Chlamydomonas) (Merchant et al. 2007; Mao et al. 2014), increasing the number of cDNA sequences for AS analysis from 0.8 million (Campbell et al. 2006) to 116 million reads (Marquez et al. 2012) led to a significant increase in the number of AS genes (11,465 vs. 5,313) and alternative splice isoforms (30,598 vs. 8,264) detected.

The degree of AS revealed by our study in *C. reinhardtii*, 19.9% of multi-exon genes, remains lower than in higher plants (Barbazuk et al. 2008; Filichkin et al. 2010; Lu et al. 2010; Marquez et al. 2012; Syed et al. 2012). For example, AS analyses in *A. thaliana* (Marquez et al. 2012) and *O. sativa* (Lu et al. 2010) revealed that ~61% and ~48% of multi-exon genes undergo AS respectively. Note however that these studies used datasets of over 100 million sequences, more than 10 times the size of ours. It is therefore expected that a large fraction of the *C. reinhardtii* AS landscape remains to be uncovered, especially by varying conditions for RNA sampling. In *C. reinhardtii* as in Arabidopsis (40%) and rice (47%), retention of intron was the prevalent mode of AS (Marquez et al. 2012; Zhang et al. 2010).

The diversity of serine/arginine-rich RNA-binding (SR) proteins is believed to underline the complexity of AS in plants (Barta et al. 2008). We therefore investigated the repertoire of SR proteins in the *C. reinhardtii* genome (see Supplemental Table 1). We identified 9 SR proteins, defined as combining one or two RRM domains followed by a serine/arginine-rich (RS) domain

(Barbosa-Morais et al. 2006; Barta et al. 2008) and 7 SR-like proteins also possibly involved in regulating splicing. Some are clearly related to the 19 SR proteins reported in *Arabidopsis*, but others are more animal-like, or altogether new. As in higher plants, AS is observed in several SR protein genes, usually leading to protein truncation.

Validation and biological significance of retained introns

To validate the PASA results, we mapped the reads to the genome and extracted the reads mapping (at least 80% identity) to each intron and to its left and right flanks. A score was computed, where a read mapping to the intron and both flanks was counted as 1.0, while a read mapping entirely within the intron was counted as 0.5 and a read mapping to the intron and a single flank was counted as 0.25. PASA Retained introns were considered validated if their score exceeded 1. This cutoff was defined empirically, by individual examination of 21 randomly selected loci using data available on the Phytozome browser and a mapping of 90 Illumina RNA-Seq experiments deposited in the SRA database and visualized with the IGV browser (Supplemental Table 3). Even though the former data is less comprehensive than ours and the latter is based on Illumina short reads that are not ideal to evaluate splicing, this was the only external data available to us for validation. Our criteria were that the retained introns should not be in a region annotated as a repeat element and that either of the following conditions were fulfilled: (i) the retained intron was present in a gene model from one of the annotations displayed on Phytozome; (ii) mapping of individual ESTs (not of the error-prone EST assembly) by Phytozome showed substantial evidence for intron retention; (iii) mapping of Illumina reads was mostly non-ambiguous and coverage over the intron was of the same order of magnitude as that over the nearby exons, with several reads spanning both junctions. This test confirmed 15

loci of AS (14 with scores >= 1, one with score=0) and invalidated 6 loci (all with scores =<1.25 except one with a score 3.75, a highly expressed gene). Using this cutoff leads to considering 1,094 (72%) of the retained introns as validated. Even if a number of false positives may be included in this list, our analysis suggests that most of the retained intron events identified by PASA are authentic and effectively increase transcriptome complexity. The 10 longest retained introns (>1300 nt) were also examined individually: 4 were confirmed as long retained introns and 2 showed no evidence for splicing and corresponded to annotation errors or pseudogenes (other cases could not be resolved because of ambiguous EST mapping).

Intron retention can be an artefact of RNA preparation, in the sense that non-spliced but poly-adenylated pre-mRNAs can be extracted along with mature cytosolic mRNAs and contribute to the library. We therefore repeated the validation step, but using only the 454 reads that showed evidence for splicing, i.e. that mapped to the genome with at least one alignment gap (see Methods for details) (3,275,045 reads, ~46%). This showed that at least 781 out of the 1,521 retained introns were undoubtedly found in spliced mRNAs. Because the 454 reads were obtained from mRNA purified on oligo-(dT) columns, we can consider that most of these transcripts were poly-adenylated and thus potentially translatable. Note that because of limited read length (median length 397 nt, compared to a median exon length of 156 nt), this more stringent analysis is bound to miss distantly spaced introns, leading to an underestimation of AS.

Yet, these intron retention events may correspond to "splicing noise", i.e. failures of the splicing machinery to unambiguously interpret the sequence signals in the pre-mRNA. These events have a different biological significance compared to true alternative splicing, selected during evolution to actually generate diversity in the mRNA and protein products generated from a single gene. To try and distinguish between the two, we analyzed the coding capacity of the retained introns. By matching the genomic coordinates of retained introns (1,521) to that of the

PASA-updated Phytozome Chre v5.3.1 gene models, we found that 1,101 (72% of the 1,521 retained introns) were flanked by entirely coding CDS exons and we termed them "CDS Retained introns"(Supplemental Figure 1). For those we searched for stop codons in the frame used by the upstream CDS exon. In the end, 315 retained introns were found to code for an additional peptide sequence within the protein, conserving the downstream sequence (they are marked in Supplemental Table 2 as "fully coding"). The other retained introns modify the sequence of the protein downstream, and can thus be expected to lead to the formation of a non-functional protein. In addition, a fraction of them lead to premature translation termination before the last exon-exon junction, which most of the time should lead to NMD. A recent genome-wide AS study in *A. thaliana* has shown that most AS isoforms are targeted to this surveillance pathway (Marquez et al. 2012). This does not mean that retained introns have no biological significance, as inactivation of protein production or production of interfering truncated proteins can be selected by evolution as means to regulate gene expression. For example, In the related alga *Volvox carteri*, the sex-regulated AS of the *MAT3* gene mostly leads to truncated versions of the Rb protein (Ferris et al. 2010).

Retained introns are less GC-rich and shorter than constitutive introns

The 1,521 retained introns have a significantly (t test, p-value = 0.0246) lower GC content than the 137,270 constitutive introns. They are also significantly shorter (Table 1). Like invertebrates (Talerico and Berget 1994; Sammeth et al. 2008) and flowering plants (Reddy 2007), *C. reinhardtii* has short introns (average 272 nt) and relatively long exons (average 368 nt). We noticed that the proportion of introns that are shorter than 250 nt, a length defined as a threshold for the intron definition splicing mechanism (Fox-Walsh et al. 2005; De Conti et al. 2013), is higher in retained introns. The length of retained introns in *C. reinhardtii* is comparable

to that in plants and animals (Michael et al. 2005; Sakabe and Souza 2007; Marquez et al. 2012; Syed et al. 2012). This phenomenon suggests that shorter introns <= 250 nt, potentially spliced by the intron definition mechanism (Talerico and Berget 1994; De Conti et al. 2013), are more likely to be retained in the case of aberrant splicing. Hence, for the following analysis of intron sequence, we concentrate on introns between 50 and 250 nt long.

Retained introns possess weaker splice site and a more G-rich Branch Point compared to constitutive introns

Our dataset thus comprised 79,228 constitutive and 1,075 retained introns. To compare splice site strength and Branch point sequences on datasets of similar size, we randomly selected 1,000 introns from both Constitutive and Retained introns. Intron splicing requires the presence of sequence signals, in particular the exon-intron boundary defining splice site sequences (5' and 3' ss) and the poly-pyrimidine tract (which in metazoans can be written $U_4Y_3U_3$, subscripts representing the nucleotide repeat number) (Black 2003). To visualize the difference between 5' and 3' ss of constitutive and retained introns in *Chlamydomonas*, we generated both frequency-based and bit-based sequence logos using weblogo (Crooks et al. 2004) of these regions (Fig.3a & 3b; see Methods for details). In addition, we used a Position Specific Scoring Matrix (PSSM) for scoring the individual introns. Using both approaches, we found that the 5'ss consensus of constitutive introns is significantly stronger than that of retained introns (Fig.3c). In particular, the 5'ss of retained introns show a non-negligible proportion of nt 'C' at position +2, which is almost never found in constitutive introns. Overall, they showed lower bit scores at all positions. The sequence in the *Chlamydomonas* spliceosomal U1 snRNA that binds the 5'ss is ACUUACCUG (Kis et al. 1993). This is complementary to the 5'ss consensus (cAG^GUGaGu) except for a U-G wobble base-pair at position +3 (underlined) which accordingly has a relatively low bit-score. We observed a slight increase in the frequency of A at this position, which may

strengthen the binding with U1 and partly compensate for the overall weaker match at other positions. However, a more detailed analysis of the energetics of 5'ss-U1 pairing site is required to definitively implicate this step as controlling intron excision/retention in *Chlamydomonas*. For the polypyrimidine tract and 3' ss, the difference was less pronounced (Fig.3), and consisted essentially in a moderate increase in the prevalence of G at pos -15 and -16 (weakening the poly-pyrimidine tract) and a decrease in that of C at the -3 position.

These results are comparable to the previous alternative splicing study in this algae, where Labadorf et al. (Labadorf et al. 2010) also reported that alternatively spliced isoforms are generally associated with weak splice site signals compared to constitutively spliced isoforms. Previous studies in vertebrates have reported similar results using both genome-wide computational analyses (Stamm et al. 2000; Zheng et al. 2005; Sakabe and Souza 2007) and wet-lab experiments on specific cases of retained introns (Sterner and Berget 1993; Dirksen et al. 1995; Romano et al. 2001; Lejeune et al. 2001).

In addition to the signals above, a branch point consensus has been described in metazoans (CURAY; the underlined A is the site where the intron 5' end is ligated) (Simpson et al. 2002). However, in land plants where the genome is relatively more AT rich (Romiguier et al. 2010; Šmarda et al. 2012) and the exon-intron structure is different compared to animals (Reddy 2007; De Conti et al. 2013), this consensus is not observed and the poly-pyrimidine tract region is richer in nucleotide U (Reddy 2001; Reddy 2007). Similarly, in *Chlamydomonas* where GC content is higher than in both plants and animals (64% vs 36% for Arabidopsis and 46% for Human), the branch point most probably does not follow the animal consensus either (Merchant et al 2007) and indeed we did not find any evidence for conservation of such a sequence upstream of 3' ss in both constitutively spliced and retained introns. Hence, instead of searching a branch point consensus by similarity to the animal sequence as has been done before, we scanned the 50

nt sequence upstream of the 3'ss for a significant sequence motif of length 5 to 8 using ELPH (http://cbcb.umd.edu/software/ELPH). The best motif is defined here as the sequence whose score is highest compared to the shuffled sequences of same nucleotide composition. For constitutive introns, we found a 6-nt long motif (Fig.4a), different from the 5-nt long motif found for the retained introns (Fig.4b). The motifs started on average at position -23 and -17 with respect to the 3' ss (-1). No fully-conserved 'A' was observed, and the best conserved A was at the last position in the motif for constitutive introns but is its center for retained introns. The difference between the two motifs suggests that determination of the branch point may be a key step in defining constitutive vs. retained introns. The branch point region is initially recognized by binding of the SF1 splicing factor (called BBP in yeast), and this protein is conserved in green algae, with two paralogs (Cre12.g553750 and Cre09.g386731 in *Chlamydomonas*).

The frequency of G- and C-triplets differs between constitutive and retained introns

In addition to the above-mentioned cis-regulatory signals, introns use additional enhancer and sometimes suppressor elements to regulate their splicing (Matlin et al. 2005; Goren et al. 2006; Reddy et al. 2012). When they are found in the introns, they are called ISE (intronic splicing enhancer) and ISS (intronic splicing suppressor). Focusing on the randomly-selected short introns, we assessed the frequency of every possible sequence pattern (3-8 nt in length) in a window of 50 nt at the start and end of the intron, using SignalSleuth (Loke et al. 2005; Shen et al. 2008). We found that the 50 nt downstream of the 5'ss are enriched in G residues (p-value < 2.2e-16; see methods for details), in particular in G triplets. This was more pronounced in constitutive than in retained introns (Fig.5a) and the difference was statistically significant (Fig.5c). McGollough and Berget (McCullough and Berget 2000) observed in short vertebrate introns that multiple G-triplets near 5' ss function as an ISE and facilitate splicing by binding to

U1snRNP. Multiple G-triplets or G-runs are a well-established ISE in vertebrate systems (Nussinov 1989; McCullough and Berget 1997) but Goren and coworkers (Goren et al. 2006) showed that such regulatory elements can have positive or negative effects depending on their context and location.

For the 50 nucleotides, upstream of the 3'ss, we observed that C-triplets were abundant for both constitutive and retained introns (Fig.5b), and again statistical tests confirmed that they were significantly enriched in constitutive introns over retained introns (Fig.5c). This is not a mere consequence of the stronger poly-pyrimidine tract, because C-triplet enrichment extends further upstream. Although this has not been found in other systems, we envision that multiple C-triplets could serve as a 3' ISE in *Chlamydomonas*.

Methods

Data

The *C. reinharditii* genome sequence and annotation model used for our analysis is Phytozome Assembly v5.3.1 (https://phytozome.jgi.doe.gov/pz/portal.html). The Sanger ESTs include those processed from 309,185 raw trace files (Liang et al. 2007) , 29,019 additional ESTs for the cDNA library and 1,112 obtained from the Chlamy Center (http://www.chlamy.org). We used pyrosequencing cDNA reads (6,317,641) from JGI (http://genome.jgi-psf.org/chlamy/chlamy.info.html) and Genoscope cDNAs (689,548) obtained from NCBI SRA (http://www.ncbi.nlm.nih.gov/sra;ERX009289 to ERX009291). The two datasets (Sanger and pyro) were processed separately to remove method-based sequence contaminants. Sanger ESTs were cleaned using a cDNA-termini identification approach (Liang et al. 2007). Both Sanger ESTs and pyrosequencing sequence reads were processed using SeqClean, a program bundled with PASA package (Haas et al. 2003; Campbell et al. 2006), which screens for sequencing

adapters, vector sequence, regions of low read-quality, short sequence fragments and regions of low complexity/repeats using DUST.

Identification of Spliced Alignments Using GMAP, BLAT and PASA

Genome-guided assembly of the dataset was performed using CLC genomics (http://www.clcbio.com/). We used an automated pipeline of PASA (Program to assemble spliced alignment) version r20140417 (Haas et al. 2003; Campbell et al. 2006) integrated with two popular mapping programs, GMAP version 2014-06-10 (Wu and Watanabe 2005) and BLAT version 35 (Kent 2002), to deduce alternatively spliced isoforms. The whole procedure followed that described in Campbell et al. (2006) with the following modifications: 1) the minimum cDNA-to-genome alignment threshold was 95% identity, 2) the minimum threshold of mapped cDNA reads was 90% coverage, 3) Three bases were required to be perfect match at splice boundary, 4) the BLAT mapping program was used to catch alignments missed by GMAP, 5) Chre v5.3.1 transcripts, both primary and alternative, were used as full length cDNAs for training PASA. After PASA cDNA assembly and spliced variant identification, the Chre v5.3.1 gene annotation was uploaded in the PASA portal to consolidate *de novo* deduced models and to update existing gene models. Integration of Chre v5.3.1 gene annotations used default settings as described on the PASA web portal (http://pasa.sourceforge.net/).

Validation of retained introns

All of the constitutively spliced introns and retained introns were extracted from the PASA database using in-house Perl scripts. The genomic coordinates of introns were used to make sure the categories of introns were non-redundant. To compute the validation score for retained introns, all the processed reads (7,201,991) were mapped to *C. reinharditii* genome using GMAP version 2014-06-10. Alignment results were obtained in gff3 format. Alignment gaps (i.e.

introns) were identified by the presence of N in CIGAR format (Li et al. 2009). Aligned reads overlapping retained introns were detected using Bioconductor packages *Biostrings* (Pages et al. 2008) and *Genomicranges* (Lawrence et al. 2013). In addition, Illumina RNA-Seq data sets retrieved from SRA (Supplemental Table 3) were aligned using bwa aln (Li et al. 2009) and visualized on the IGV browser (Robinson et al. 2011). All the codes used for processing will be made available upon request.

Splice site scores

The splice site score for the 5' site was calculated using the 9 nucleotide positions that flank the junction, from -2 to +6 with position 1 being the first nucleotide of the intron (see Fig.3). For the 3' splice site, we chose 24 nucleotides as described in (Tian et al. 2007), i.e., 22 nucleotides upstream of intron 3' boundary and 2 nucleotides from the following exon (Fig.3). A Perl script was used to generate position specific scoring matrices (PSSM) for the 9 nt and 24 nt regions respectively, with matrix elements given per the equation 1.

$$m_{ij} = \log2(f_{ij}/g_i) \quad (1)$$

Here $m_{i,j}$ is the score of nucleotide i at position j in the *matrix*, f_{ij} is the frequency of nucleotide i at position j, and g_i is the frequency of nucleotide i in the whole region. The score for each individual sequence was calculated by equation 2.

$$Score = \Sigma_j m_{i,j} \quad (2)$$

Nucleotide Profiling

To examine nucleotide profile differences, we took two windows of size 50 nt from both splice sites immediately adjacent to exons, that is 50 nt in introns downstream of the 5'ss and 50 nt in introns upstream of the 3'ss respectively. The SignalSleuth package (Zhao et al. 2014) was

used to calculate the frequency of every possible pattern of nucleotide with a length from 2 to 8 nt within the 50 nt window. The output of SignalSleuth includes a matrix file for every possible pattern, filtered and sorted based on their frequency compared to the background. The patterns that are consistently more frequent than other patterns over the window are considered for further analyses, including line graph comparisons and significance testing. Single nucleotide enrichment in the 50 nt window were assessed using t-test between all pairs of single nt probabilities among the sequences.

Statistical analyses

PSSM scores and motif frequency distributions were visualized using box plots and differences in medians were statistically assessed within R, (http://www.r-project.org/) using a Wilcoxon test with 95% level of confidence. In Wilcoxon test, the alternate hypothesis chosen was that length of Constitutive introns is higher than the Retained introns with the null hypothesis being they are equal. Non-parametric Wilcoxon tests were used because PSSM scores and frequency values are not normally distributed as required for a Student's t-test. For GC content, in which the data was normally distributed, the t-test was used.

List of abbreviations

5'ss – 5' splice site
3'ss – 3' splice site
CDS – Coding Sequence
ESE – exonic splicing enhancer
ESS – exonic splicing silencer
GFF3 - General Feature Format
ISE – intronic splicing enhancer
ISS – intronic splicing silencer

nt – nucleotide
PSSM –Position Specific Scoring Matrix
NMD – non-mediated decay
AS – Alternative splicing
PASA - Program to Assemble Spliced Alignment

Authors' contributions

CL managed and coordinated the project. PKRK and CL conceived the study. PKRK carried out implementation and drafted the manuscript. OV conducted SR protein analysis and individual intron inspection and helped devise validation tests. All authors participated in manuscript writing.

Acknowledgement

Funding for this project is provided by a grant award from the Ohio Plant Biotechnology Consortium to CL. This work was partially supported by the National Institutes of Health [1R15GM94732-1 A1 to CL] and Centre National de la Recherche Scientifique [ANR-11-LABX-0011-DYNAMO to OV]. The authors thank Mario Stanke, Lin Liu and Trey Moler for their participation in this project and Marina Cavaioulo for giving access to her mapping of Illumina reads.

References

Barbazuk W, Fu Y, McGinnis K (2008) Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. Genome Research 18:1381–1392. doi: 10.1101/gr.053678.106

Barbosa-Morais NL, Carmo-Fonseca M, Aparício S (2006) Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res 16:66–77. doi: 10.1101/gr.3936206

Barta A, Kalyna M, Lorković ZJ (2008) Plant SR Proteins and Their Functions. In: Reddy ASN, Golovkin M (eds) Nuclear pre-mRNA Processing in Plants. Springer Berlin Heidelberg, pp 83–102

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414.

Blaby IK, Blaby-Haas CE, Tourasse N, Hom EFY, Lopez D, Aksoy M, Grossman A, Umen J, Dutcher S, Porter M, King S, Witman GB, Stanke M, Harris EH, Goodstein D, Grimwood J, Schmutz J, Vallon O, Merchant SS, Prochnik S (2014) The Chlamydomonas genome project: a decade on. Trends in Plant Science. doi: 10.1016/j.tplants.2014.05.008

Black D (2003) mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry 72:291–336. doi: 10.1146/annurev.biochem.72.121801.161720

Campbell M, Haas B, Hamilton J, Mount S, Buell C (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 7:327. doi: 10.1186/1471-2164-7-327

Carvalho RF, Feijão CV, Duque P (2013) On the physiological significance of alternative splicing events in higher plants. Protoplasma 250:639–650. doi: 10.1007/s00709-012-0448-9

Colak D, Ji S-J, Porse BT, Jaffrey SR (2013) Regulation of Axon Guidance by Compartmentalized Nonsense-Mediated mRNA Decay. Cell 153:1252–1265. doi: 10.1016/j.cell.2013.04.056

Crooks G, Hon G, Chandonia J, Brenner S (2004) WebLogo: a sequence logo generator. Genome research 14:1188–1190. doi: 10.1101/gr.849004

De Conti L, Baralle M, Buratti E (2013) Exon and intron definition in pre-mRNA splicing. WIREs RNA 4:49–60. doi: 10.1002/wrna.1140

Dirksen WP, Sun Q, Rottman FM (1995) Multiple Splicing Signals Control Alternative Intron Retention of Bovine Growth Hormone Pre-mRNA. J Biol Chem 270:5346–5352. doi: 10.1074/jbc.270.10.5346

Falciatore A, Merendino L, Barneche F, Ceol M, Meskauskiene R, Apel K, Rochaix J (2005) The FLP proteins act as regulators of chlorophyll synthesis in response to light and plastid signals in Chlamydomonas. Genes & development 19:176–187.

Ferris P, Olson BJSC, Hoff PLD, Douglass S, Casero D, Prochnik S, Geng S, Rai R, Grimwood J, Schmutz J, Nishii I, Hamaji T, Nozaki H, Pellegrini M, Umen JG (2010) Evolution of an Expanded Sex-Determining Locus in Volvox. Science 328:351–354. doi: 10.1126/science.1186222

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong W-K, Mockler TC (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res 20:45–58. doi: 10.1101/gr.093302.109

Fox-Walsh KL, Dou Y, Lam BJ, Hung S, Baldi PF, Hertel KJ (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. PNAS 102:16176–16181. doi: 10.1073/pnas.0508489102

Fukuzawa H, Miura K, Ishizaki K, Kucho K, Saito T, Kohinata T, Ohyama K (2001) Ccm1, a regulatory gene controlling the induction of a carbon-concentrating mechanism in Chlamydomonas reinhardtii by sensing CO2 availability. PNAS 98:5347–5352. doi: 10.1073/pnas.081593498

Ge Y, Porse BT (2014) The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. BioEssays 36:236–243. doi: 10.1002/bies.201300156

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G (2006) Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers. Molecular Cell 22:769–781. doi: 10.1016/j.molcel.2006.05.008

Guo S, Zheng Y, Joung J-G, Liu S, Zhang Z, Crasta OR, Sobral BW, Xu Y, Huang S, Fei Z (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. BMC Genomics 11:384. doi: 10.1186/1471-2164-11-384

Haas BJ, Delcher AL, Mount SM, Wortman JR, Jr RKS, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucl Acids Res 31:5654–5666. doi: 10.1093/nar/gkg770

Kalyna M, Lopato S, Voronin V, Barta A (2006) Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. Nucl Acids Res 34:4395–4405. doi: 10.1093/nar/gkl570

Kent W (2002) BLAT - The BLAST-Like Alignment Tool. Genome Research 12:656–664.

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11:345–355. doi: 10.1038/nrg2776

Khodor YL, Menet JS, Tolan M, Rosbash M (2012) Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. RNA 18:2174–2186. doi: 10.1261/rna.034090.112

Kis M, Jakab G, Pollak T, Branlant C, Solymosy F (1993) Nucleotide sequence of U1 RNA from a green alga, Chlamydomonas reinhardtii. Nucleic Acids Res 21:2255.

Labadorf A, Link A, Rogers M, Thomas J, Reddy A, Ben-Hur A (2010) Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. BMC Genomics 11:114. doi: 10.1186/1471-2164-11-114

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9:e1003118. doi: 10.1371/journal.pcbi.1003118

Lejeune F, Cavaloc Y, Stevenin J (2001) Alternative Splicing of Intron 3 of the Serine/Arginine-rich Protein 9G8 Gene IDENTIFICATION OF FLANKING EXONIC SPLICING ENHANCERS AND INVOLVEMENT OF 9G8 AS A TRANS-ACTING FACTOR. J Biol Chem 276:7850–7858. doi: 10.1074/jbc.M009510200

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi: 10.1093/bioinformatics/btp352

Li JB, Lin S, Jia H, Wu H, Roe BA, Kulp D, Stormo GD, Dutcher SK (2003) Analysis of Chlamydomonas reinhardtii genome structure using large-scale sequencing of regions on linkage groups I and III. J Eukaryot Microbiol 50:145–155.

Liang C, Wang G, Liu L, Ji G, Liu Y, Chen J, Webb JS, Reese G, Dean JFD (2007) WebTraceMiner: a web service for processing and mining EST sequence trace files. Nucleic Acids Res 35:W137–W142. doi: 10.1093/nar/gkm299

Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. Proceedings of the National Academy of Sciences 98:11193–11198. doi: 10.1073/pnas.201407298

Loke JC, Stahlberg EA, Strenski DG, Haas BJ, Wood PC, Li QQ (2005) Compilation of mRNA Polyadenylation Signals in Arabidopsis Revealed a New Signal Element and Potential Secondary Structures. Plant Physiol 138:1457–1468. doi: 10.1104/pp.105.060541

Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W, Huang X, Han B (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res 20:1238–1249. doi: 10.1101/gr.106120.110

Macknight R, Duroux M, Laurie R, Dijkwel P, Simpson G, Dean C (2002) Functional Significance of the Alternative Transcript Processing of the Arabidopsis Floral Promoter FCA. Plant Cell 14:877–888. doi: 10.1105/tpc.010456

Mao R, Raj Kumar PK, Guo C, Zhang Y, Liang C (2014) Comparative Analyses between Retained Introns and Constitutively Spliced Introns in Arabidopsis thaliana Using Random Forest and Support Vector Machine. PLoS ONE 9:e104049. doi: 10.1371/journal.pone.0104049

Marquez Y, Brown J, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res 22:1184–1279. doi: 10.1101/gr.134106.111

Matlin AJ, Clark F, Smith CWJ (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6:386–398. doi: 10.1038/nrm1645

McCullough AJ, Berget SM (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol 17:4562–4571.

McCullough AJ, Berget SM (2000) An Intronic Splicing Enhancer Binds U1 snRNPs To Enhance Splicing and Select 5? Splice Sites. Mol Cell Biol 20:9225–9235.

Merchant S, Prochnik S, Vallon O, Harris E, Karpowicz S, Witman G, Terry A, Salamov A, Fritz-Laylin L, Marechal-Drouard L, other s (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318:245. doi: 10.1126/science.1143609

Michael IP, Kurlender L, Memari N, Yousef GM, Du D, Grass L, Stephan C, Jung K, Diamandis EP (2005) Intron Retention: A Common Splicing Event within the Human Kallikrein Gene Family. Clinical Chemistry 51:506–515. doi: 10.1373/clinchem.2004.042341

Moseley JL, Page MD, Alder NP, Eriksson M, Quinn J, Soto F, Theg SM, Hippler M, Merchant S (2002) Reciprocal Expression of Two Candidate Di-Iron Enzymes Affecting Photosystem I and Light-Harvesting Complex Accumulation. Plant Cell 14:673–688. doi: 10.1105/tpc.010420

Nussinov R (1989) Conserved Signals Around the 5′ Splice Sites in Eukaryotic Nuclear Precursor mRNAs: G-Runs are Frequent in the Introns and C in the Exons Near Both 5′ and 3′ Splice Sites. Journal of Biomolecular Structure and Dynamics 6:985–1000. doi: 10.1080/07391102.1989.10506526

Pages H, Gentleman R, Aboyoun P, DebRoy S (2008) Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2:160.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415. doi: 10.1038/ng.259

Quesada V, Macknight R, Dean C, Simpson GG (2003) Autoregulation of FCA pre-mRNA processing controls Arabidopsis flowering time. The EMBO Journal 22:3142–3152. doi: 10.1093/emboj/cdg305

Reddy A (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol 58:267–361. doi: 10.1146/annurev.arplant.58.032806.103754

Reddy A (2001) Nuclear Pre-mRNA Splicing in Plants. Crit. Rev. Plant Sci 20:523–571.

Reddy ASN, Rogers MF, Richardson DN, Hamilton M, Ben-Hur A (2012) Deciphering the Plant Splicing Code: Experimental and Computational Approaches for Predicting Alternative Splicing and Splicing Regulatory Elements. Front Plant Sci. doi: 10.3389/fpls.2012.00018

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative Genomics Viewer. Nat Biotechnol 29:24–26. doi: 10.1038/nbt.1754

Romano M, Marcucci R, Baralle FE (2001) Splicing of constitutive upstream introns is essential for the recognition of intra-exonic suboptimal splice sites in the thrombopoietin gene. Nucl Acids Res 29:886–894. doi: 10.1093/nar/29.4.886

Romiguier J, Ranwez V, Douzery EJP, Galtier N (2010) Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. Genome Res 20:1001–1009. doi: 10.1101/gr.104372.109

Sakabe NJ, Souza SJ de (2007) Sequence features responsible for intron retention in human. BMC Genomics 8:59. doi: 10.1186/1471-2164-8-59

Sakharkar MK, Perumal BS, Sakharkar KR, Kangueane P (2005) An Analysis on Gene Architecture in Human and Mouse Genomes. In Silico Biology 5:347–365.

Sammeth M, Foissac S, Guigó R (2008) A General Definition and Nomenclature for Alternative Splicing Events. PLoS Comput Biol 4:e1000147. doi: 10.1371/journal.pcbi.1000147

Seo PJ, Kim MJ, Ryu J-Y, Jeong E-Y, Park C-M (2011) Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. Nat Commun 2:303. doi: 10.1038/ncomms1303

Shen Y, Ji G, Haas BJ, Wu X, Zheng J, Reese GJ, Li QQ (2008) Genome level analysis of rice mRNA 3?-end processing signals and alternative polyadenylation. Nucleic Acids Res 36:3150–3161. doi: 10.1093/nar/gkn158

Šmarda P, Bureš P, Šmerda J, Horová L (2012) Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. New Phytologist 193:513–521. doi: 10.1111/j.1469-8137.2011.03942.x

Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ (2000) An Alternative-Exon Database and Its Statistical Analysis. DNA and Cell Biology 19:739–756. doi: 10.1089/104454900750058107

Sterner DA, Berget SM (1993) In vivo recognition of a vertebrate mini-exon as an exon-intron-exon unit. Mol Cell Biol 13:2677–2687. doi: 10.1128/MCB.13.5.2677

Syed NH, Kalyna M, Marquez Y, Barta A, Brown JWS (2012) Alternative splicing in plants - coming of age. Trends Plant Sci 17:616–623. doi: 10.1016/j.tplants.2012.06.001

Talerico M, Berget SM (1994) Intron definition in splicing of small Drosophila introns. Molecular and Cellular Biology 14:3434–3445. doi: 10.1128/MCB.14.5.3434

Tian B, Pan Z, Lee JY (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res 17:156–165. doi: 10.1101/gr.5532707

Wang B, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. PNAS 103:7175–7180. doi: 10.1073/pnas.0602039103

Wong JJ-L, Ritchie W, Ebner OA, Selbach M, Wong JWH, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, Thoeng A, Khoo T-L, Bailey CG, Holst J, Rasko JEJ (2013) Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation. Cell 154:583–595. doi: 10.1016/j.cell.2013.06.052

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875. doi: 10.1093/bioinformatics/bti310

Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L, Tian W, Tao Y, Kristiansen K, Zhang X, Li S, Yang H, Wang J, Wang J (2010) Deep RNA

sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 20:646–654. doi: 10.1101/gr.100677.109

Zhang MQ (1998) Statistical Features of Human Exons and Their Flanking Regions. Hum Mol Genet 7:919–932. doi: 10.1093/hmg/7.5.919

Zhao Z, Wu X, Kumar PKR, Dong M, Ji G, Li QQ, Liang C (2014) Bioinformatics Analysis of Alternative Polyadenylation in Green Alga Chlamydomonas reinhardtii Using Transcriptome Sequences from Three Different Sequencing Platforms. G3 (Bethesda) 4:871–883. doi: 10.1534/g3.114.010249

Zheng CL, Fu X-D, Gribskov M (2005) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. RNA 11:1777–1787. doi: 10.1261/rna.2660805

Figure Legends

Fig.1. Type and extent of alternatively spliced genes and events in *Chlamydomonas reinhardtii*

a. An example of a splice isoform with intron retention as viewed in PASA web site. Splice sites are shown in light blue, exons in black, coding regions as predicted by PASA in red, and spliced out introns by the thin connecting lines.
b. Percentage of AS genes displaying the 5 major AS types. The same genes may be assigned to multiple types. skipped_exon = an exon that has been skipped in at least one isoform; retained_intron = an intron that has been spliced out in at least one isoform; alternate_exon = isoforms with mutually exclusive exons; alt_donor = isoforms with alternative donor sites (5' splice site); alt_acceptor = isoforms with alternative acceptor sites (3' splice site). Please note that the numbers don't add up to 100% as there are more than one AS event per gene.
c. Proportion of AS events.

Fig.2. Number of splice sites at the alternative location in isoforms.
Number of splice sites at the alternative position. Delta in the x-axis denotes the number of bases different between splice sites in splice isoforms. Interactive figure with varying delta and to see underlying text data can be accessed at the PASA website [http://bioinfolab.miamioh.edu/cgi-bin/PASA_r20140417/cgi-bin/alt_donor_acceptor_deltas.cgi?db=Chre_AS&bin_size=1&max_delta=100]

Fig.3. Comparison of splice site motifs for short (<= 250 nt) constitutive and retained introns.

a. Frequency and information bit logos for constitutive introns.

b. Frequency and information bit logos for retained introns.

c. Score distribution of 5' and 3' splice site of short (<=250 nt) constitutive and retained introns. The PSSM for all intron splice sites was used to score individual introns. The median score of Constitutive introns is significantly greater than that of Retained introns by a Wilcoxon test [p-value < 2.2e-16 and < 1.16e-15, respectively].

Fig.4. Frequency and information bit logos of motifs found in the last 50 nt window in constitutive (a) and retained (b) introns.

Fig.5. G-triplets profile within 50 nt downstream of the 5' splice site and C-triplets profile within 50 nt upstream of 3' splice site.

a. Profile of G-triplets at every position within short (<=250 nt) constitutive and retained introns, 50 nt downstream of 5'ss. The scan for patterns was performed from left to right, e.g.,the sequence (ggggg) contains 3 G-triplets at positions 1,2,3.

b. Profile of C-triplets at every position within short (<=250 nt) constitutive and retained introns, 50 nt upstream of 3'ss .

c. Quantitative comparison of G- and C-triplets abundance between constitutive and retained introns over the first and last 50nt windows, respectively. The median frequencies for constitutive introns are significantly greater than for retained introns, with p-value = 2.67e-9 and 2.07e-05 respectively.

Supplemental Fig.1. Example of alternative splicing event supported only by Augustus annotation.

a. Splicing graph of the gene structure as viewed in PASA website [http://bioinfolab.miamioh.edu/cgi-bin/PASA_r20140417/cgi-bin/assembly_alt_splice_info.cgi?db=Chre_AS&cdna_acc=asmbl_94&SHOW_ALL=1&SHOW_ALIGNMENTS].

b. and **c.** show alternative splicing forms "alternative acceptor" and "retained intron" as inferred from comparing two isoforms called asmbl_94 and asmbl_95. Splice sites are shown in light blue, exons in black, coding regions as predicted by PASA in red, and spliced out introns by the thin connecting lines. Highlighted purple box denote the change between isoforms. Both isoforms asmbl_94 and asmbl_95 are supported by only Augustus annotations "Cre01.g003100.t1.3" and "Cre01.g003100.t2.1" respectively. An intron retained between the coding exons is denoted as CDS retained intron.

Tables

Table 1: **Length distribution and GC content of constitutive and retained introns.* denotes the percentage of introns with length less than or equal to 250 nt.**

Supplemental Table 1: *Chlamydomonas* SR proteins.

SR protein family assignments were based on (Barbosa-Morais et al. 2006; Barta et al. 2008). RRM: RNA recognition motif; RS: arginine/serine-rich region; ZnK: Zinc knuckle domain; SWAP/Surp: suppressor-of-white-apricot domain; PWI: Pro-Trp-Ile domain. Parentheses within a "Known Domains" field entry indicate a weak domain feature. Au9 gene models, annotations, and corresponding browser links showing PASA transcript models are provided in the right column.

Supplemental Table 2: Retained introns and their validation scores.

Retained intron found in the PASA model with their genomic coordinates, validation scores and results of individual inspection where available.

Supplemental Table 3: SRA accessions for data.

Key Message

We report that higher proportion of genes is alternatively spliced in *Chlamydomonas reinhardtii* than previously reported and describe the sequence features responsible for major AS event retention of introns.

Authors' contributions

CL managed and coordinated the project. PKRK and CL conceived the study. PKRK carried out implementation and drafted the manuscript. OV conducted SR protein analysis and individual intron inspection and helped devise validation tests. All authors participated in manuscript writing.

CL : Chun Liang

PKRK: Praveen Kumar Raj Kumar

OV: Olivier Vallon

Figure1

Click here to download Colour figure Figure_1.tif

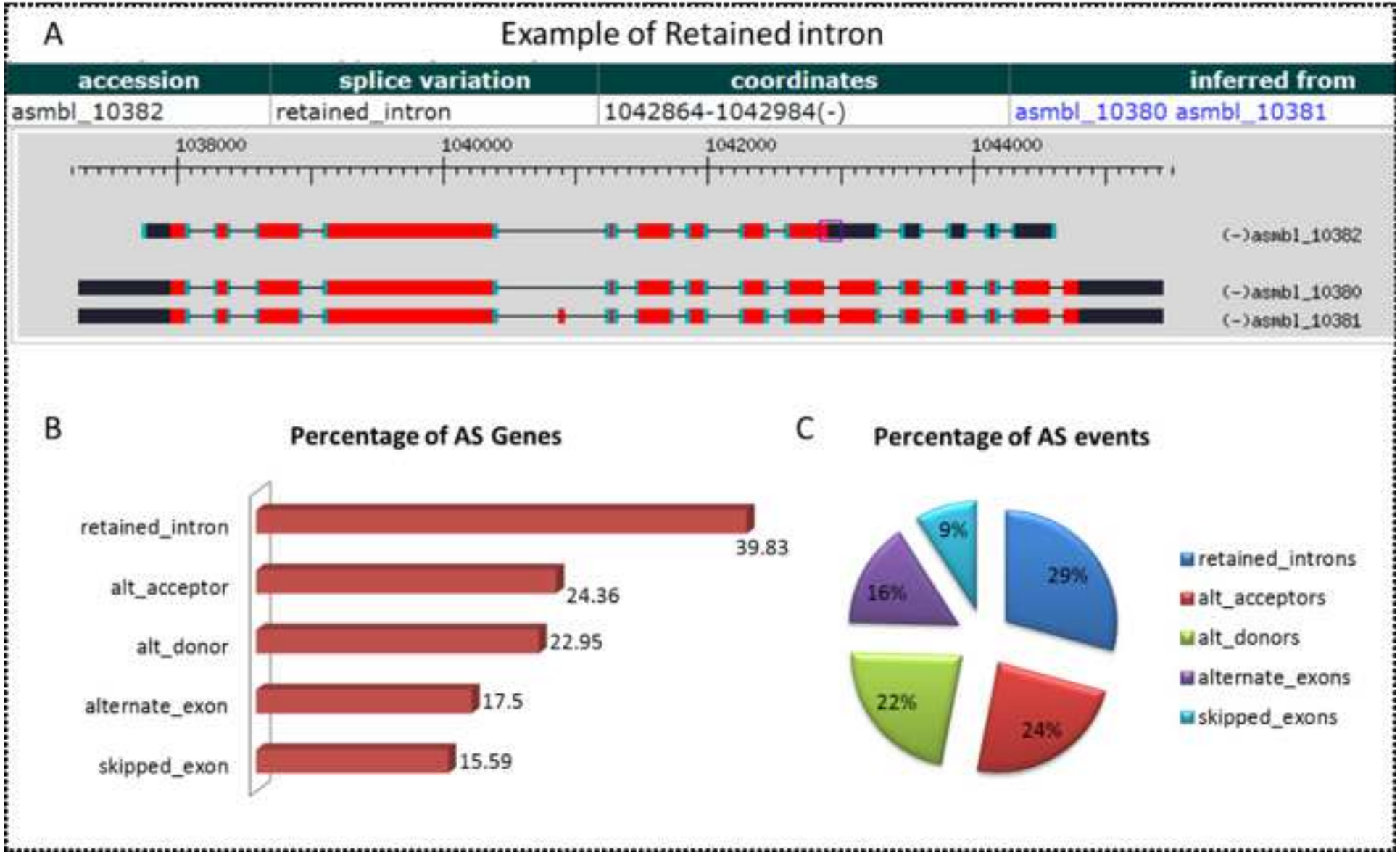

Figure2

Click here to download Colour figure Figure_2.png

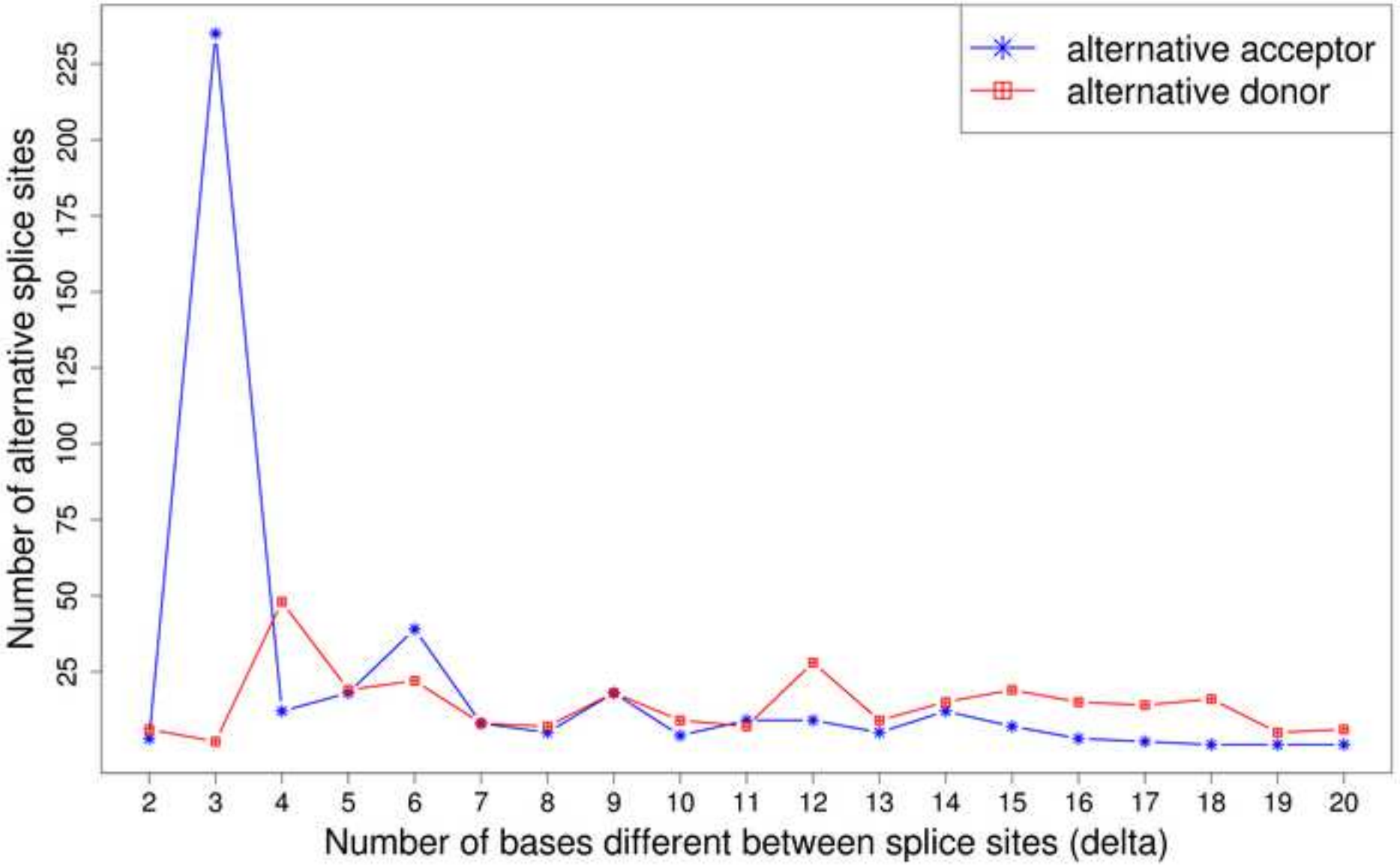

Figure3

Click here to download Colour figure Figure_3.png

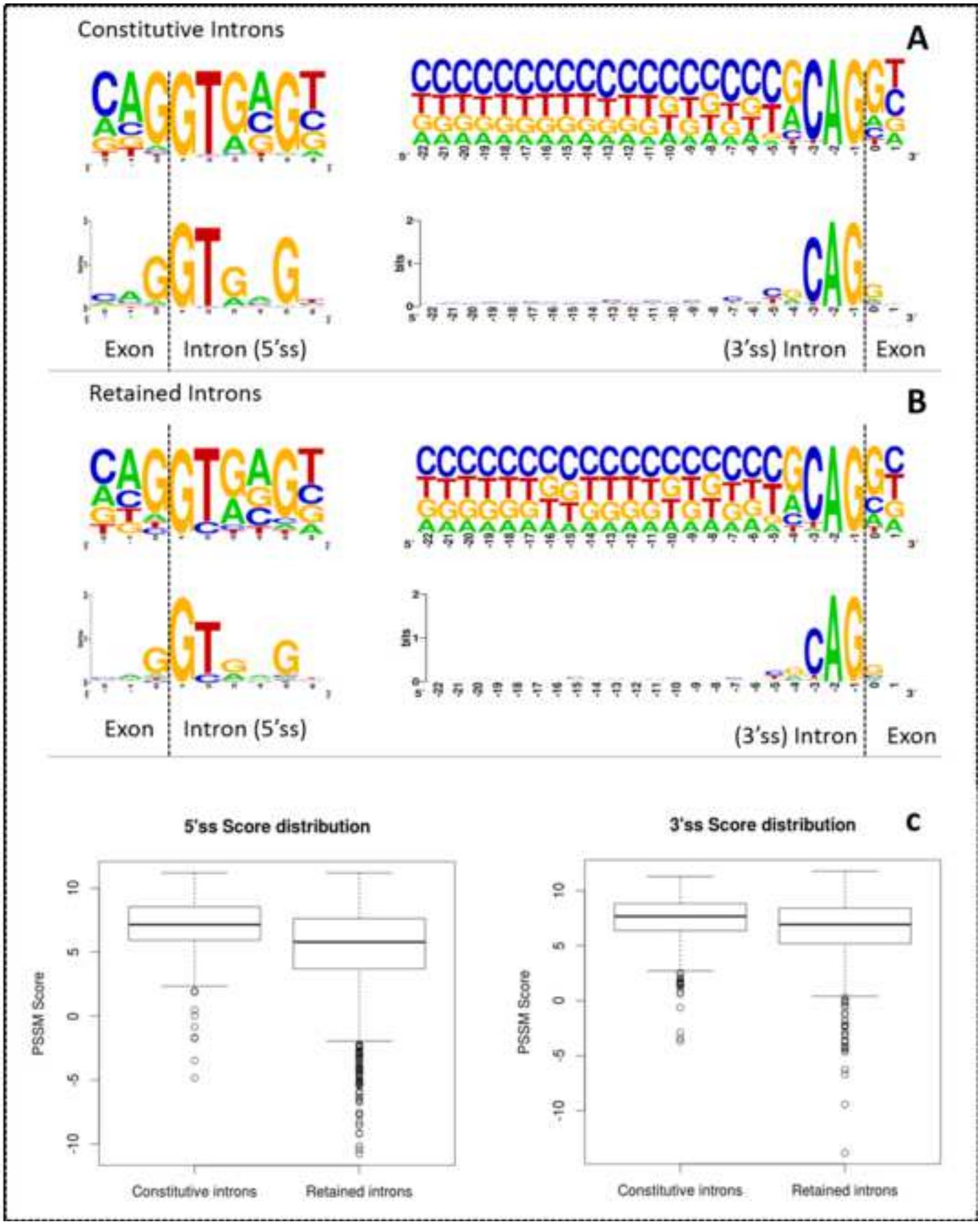

Figure4

Click here to download Colour figure Figure_4.tif

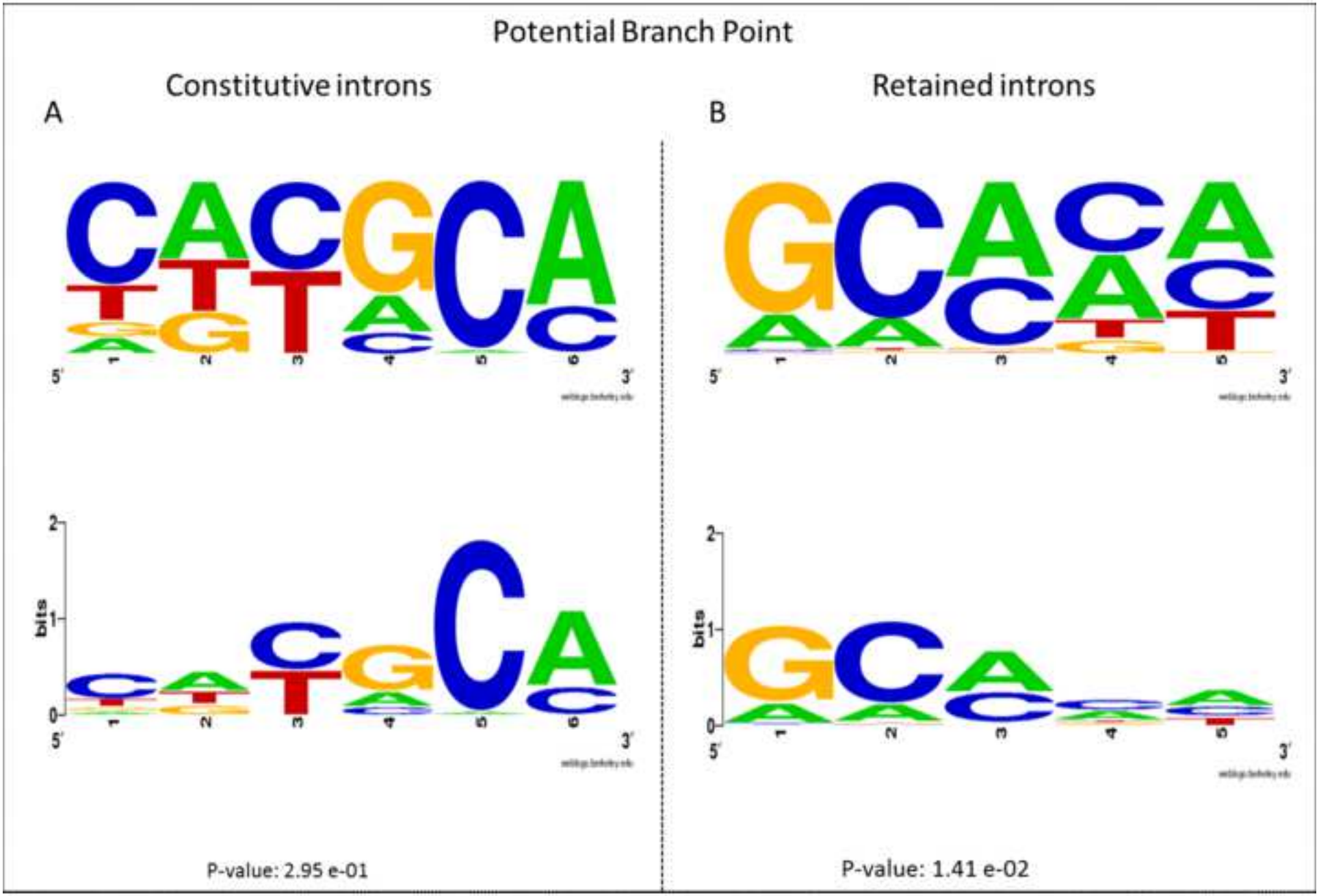

Figure5

Click here to download Colour figure Figure_5.tif

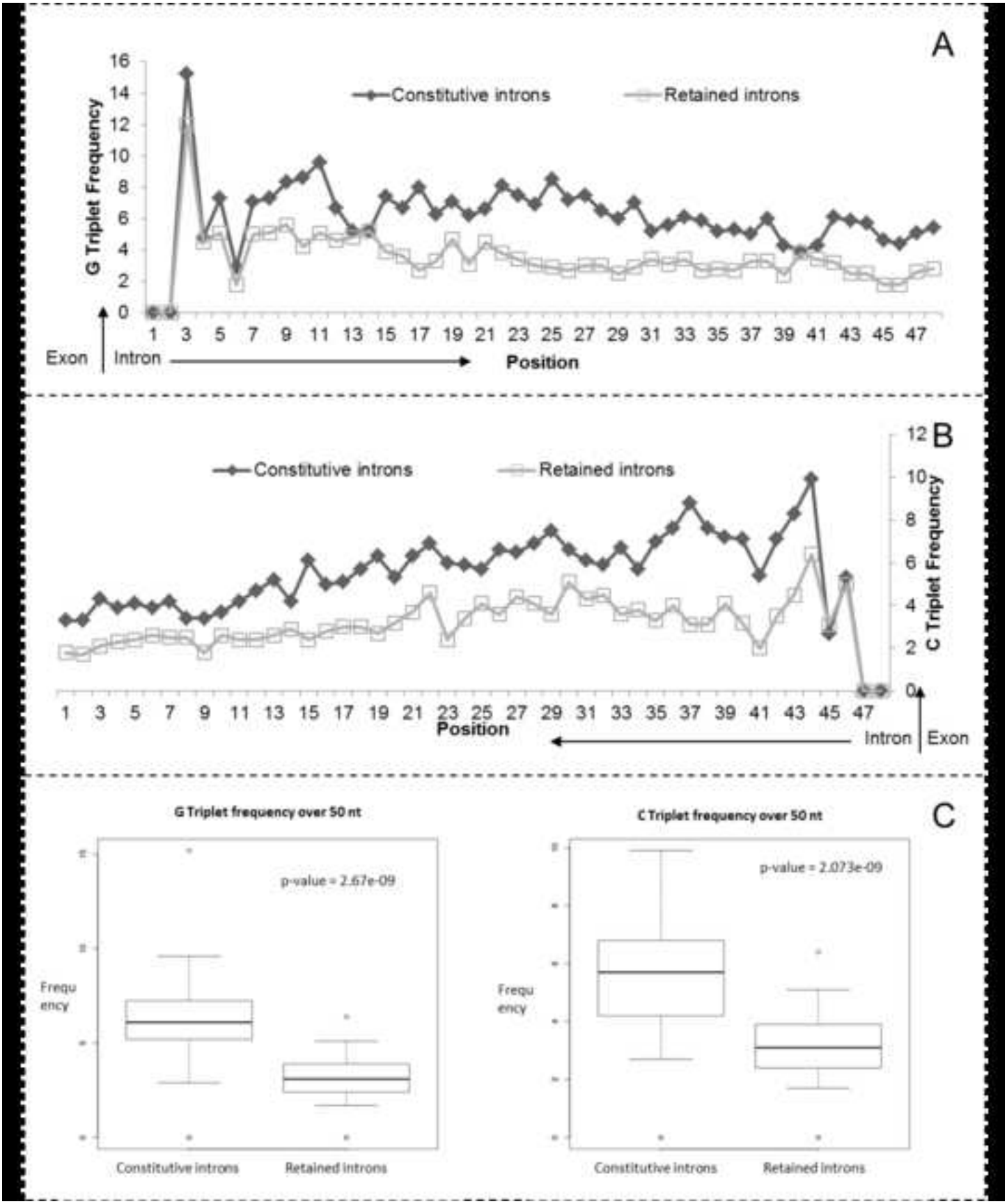

Table 1: Length distribution and GC content of constitutive and retained introns.

Intron Category	Number of introns	Mean (Median) in nt	Number of introns (length <= 250 nt)*	GC content (%)
Constitutive introns	137,270	319.87 (228)	79,228 (57.71%)	61.46
Retained introns	1,521	229.38(181)	1,075 (70.67%)	60.42