Genomic and Transcriptomic Analysis Reveals Spliced Leader Trans-Splicing in Cryptomonads

Scott William Roy*

Department of Biology, San Francisco State University, San Francisco, CA

*Corresponding author: E-mail: scottwroy@gmail.com.

Accepted: February 2, 2017

Abstract

Spliced leader trans-splicing (SLTS) is a poorly understood mechanism that is found in a diversity of eukaryotic lineages. In SLTS, a short RNA sequence is added near the 5′ ends of the transcripts of protein-coding genes by a modified spliceosomal reaction. Available data suggest that SLTS has evolved many times, and might be more likely to evolve in animals. That SLTS might be more likely to evolve in the context of the generally complex transcriptomes characteristic of animals suggests the possibility that SLTS functions in gene regulation or transcriptome diversification, however no general novel function for SLTS is known. Here, I report SLTS in a lineage of cellularly complex unicellular eukaryotes. Cryptomonads are a group of eukaryotic algae that acquired photosynthetic capacity by secondary endosymbiosis of a red alga, and that retain a reduced copy of the nucleus of the engulfed alga. I estimate that at least one-fifth of genes in the model cryptomonad *Guillardia theta* and its relative *Hanusia phi* undergo SLTS. I show that hundreds of genes in *G. theta* generate alternative transcripts by SLTS at alternative sites, however I find little evidence for alternative protein production by alternative SLTS splicing. Interestingly, I find no evidence for substantial operon structure in the *G. theta* genome, in contrast to previous findings in other lineages with SLTS. These results extend SLTS to another major group of eukaryotes, and heighten the mystery of the evolution of SLTS and its association with cellular and transcriptomic complexity.

Key words: protist molecular biology, genome evolution, gene expression.

Introduction

The transcriptomic era has brought a variety of surprises in terms of the mechanisms and complexity of gene expression and genome transcription, with organisms such as animals showing high incidences of mechanisms from alternative splicing, alternative promoter usage, alternative polyadenylation and RNA editing to transcription of diverse intergenic loci and regulatory RNAs (Lewis et al. 2003; Nilsen and Graveley 2010; Shepard et al. 2011; Peng et al. 2012; Saito et al. 2013). One poorly understood transcriptomic phenomenon is spliced leader trans-splicing (SLTS), in which abundant, short, largely homogeneous RNAs are transcribed from a relatively small number of genomic loci and added near the 5′ ends of various pre-mRNAs by a modified spliceosomal reaction (Lasda and Blumenthal 2011; Bitar et al. 2013). SLTS has been reported in several distantly-related protist lineages—including euglenids, kinetoplastids, dinoflagellates and cercozoans—as well as over one dozen animal lineages including flatworms, hydra, ctenophores, rotifers, urochordates, nematodes, copepods, amphipods, sponges and chaetognaths (Sutton and Boothroyd 1986; Krause and Hirsh 1987; Ganot et al. 2004; Zhang et al. 2007; Douris et al. 2010; Yang et al. 2015; Nowack et al. 2016).

Transcriptomic studies have begun to reveal genome-wide patterns of SLTS in various species. Transcriptomic studies in some species have revealed a high degree of alternative SLTS, in which alternative transcript isoforms of the same gene are produced by addition of SLs at alternative sites (thus each transcript isoform has an SL added at one out of two or more sites (Horiuchi and Aigaki 2006; Nilsson et al. 2010). For instance, at least 19% of SLTS genes in *Trypanosoma brucei* were found to undergo alternative SLTS (Nilsson et al. 2010). In addition, various studies have shown a close association of SLTS with the presence of multi-gene operons, with polycistronic pre-mRNA transcripts being resolved into individual monocistronic transcripts by addition of SL sequences upstream of each translation start codon (Johnson et al. 1987; Zorio et al. 1994), a process that is often coupled to polyadenylation of upstream sequences (LeBowitz et al. 1993; Kuersten et al. 1997; Jäger et al. 2007).

The peculiar phylogenetic distribution of lineages known to undergo SLTS strongly suggests recurrent independent evolution. Interestingly, SLTS is inferred to have evolved >10 times within animals compared to only a few known times in all other lineages combined (Douris et al. 2010). This pattern of recurrent evolution of SLTS with particular concentration in the most organismally and molecularly complex eukaryotic group, seems to suggest that SLTS may impart important functions. A variety of functions have been proposed or demonstrated (see Lasda and Blumenthal 2011 for a review), including trimming off of "extra" 5′ sequence from RNAs, providing a 5′ methyl cap for mRNA transcripts, providing ATG translation start codons to transcripts of genes whose genomic copies lack them, resolving polycistronic pre-mRNA transcripts into single-ORF mRNAs, and increasing regulatory control and proteomic diversity. However, many of these proposed functions (providing mRNAs with 5′ methyl caps, 5′-UTRs, translation start codons) involve producing core features of eukaryotic transcripts, features which are produced by non-SLTS species. That is, these functions appear to be redundant with presumably ancestral functions, thus, it is not clear how selection for these functions could have driven the initial origins of SLTS. Resolution of polycistronic transcriptions also seems unlikely to explain the origins of SLTS since standard translation initiation in eukaryotes is thought to limit the evolution of polycistronic gene expression in the absence of SLTS (Kozak 1999); thus the association of operons and SLTS seems more likely to be due to SLTS facilitating the evolution of operons than vice versa. For other functions, including additional levels of gene regulation and increased proteomic diversity, the evolution of the additional function for a particular gene presupposes the existence of an SLTS system, and thus also seems unlikely to explain the initial origins of SLTS within a lineage. Another possibility is that an SLTS system could evolve in a limited context (for instance, by chance, or by providing 5′ methyl caps for a small number of genes transcribed by polymerase I; Lee and Van der Ploeg 1997; Günzl et al. 2003); SLTS could then spread to other genes through random evolution of SLTS splice sites followed by degradation of the genomically encoded 5′-UTR, rendering SLTS necessary for proper gene function (Maeso et al. 2012). The relative contributions of these various possibilities to the initial evolution of SLTS within a lineage remain largely untested, thus the origins of SLTS remain mysterious.

Recently, we reported the genomic sequence of species from two lineages of unicellular eukaryotes with particularly complex subcellular compartmentalization (Curtis et al. 2012). Cryptomonads and chlorarachniophytes are groups of algae that independently acquired photosynthetic organelles by engulfing primary eukaryotic algae (red and green algae, respectively). Atypically among secondary/tertiary algae, these lineages retain as an organelle a remnant of the nucleus of the engulfed algae, called the nucleomorph, which retains its own greatly reduced genome (see Moore and Archibald 2009 for a discussion). Interestingly, transcriptomic analysis of the two species revealed a high level of transcriptomic complexity for the sequenced chlorarachniophyte, *Bigelowiella natans*, which exhibited a degree of alternative splicing unprecedented among characterized unicellular organisms (Curtis et al. 2012). Some of these alternative splicing events in *B. natans* were predicted to give rise to alternative protein isoforms with different subcellular targeting signaling, suggesting a potential role for transcriptomic complexity in the functions of these structurally complex cells (Curtis et al. 2012). No similar transcriptomic complexity was noted in the sequenced cryptophyte, *Guillardia theta*.

Transcriptomic and Genomic Evidence for SLTS in Cryptomonads

While studying the transcriptome of *G. theta* previously assembled by the MMETSP consortium (Keeling et al. 2014), I noticed that many unrelated transcripts had a shared sequence, 5′-CTCGGCGGCTGGTCAAG-3′, near their ends (fig. 1*A*). BLASTN searches of several transcripts against the genome showed that this sequence was not genomically encoded near the downstream sequence, suggesting the possibility that this sequence was a spliced leader sequence (fig. 1*A*). Consistent with this, scrutiny of the genome revealed that genic copies of the downstream (nonshared) sequences were almost universally preceded by a potential splice acceptor sequence (an AG dinucleotide; fig. 1*A*). Searches of the genome revealed 20 copies of the candidate SL sequence, each with a downstream potential donor GT dinucleotide (fig. 1*B*). Consistent with lack of observation of variant SL sequences in the transcriptomic data, no variation in the mature SL sequence was observed among putative genomic SL gene sequences. Alignment of these 20 copies showed clear similarity starting abruptly at the beginning of this sequence and extending through 23 nucleotides past the putative splice site (fig. 1*B*). All 20 copies ended in a U-rich motif flanked by purines, a motif characteristic of known SM binding sites, which is consistent with spliceosomal spliced leader trans-splicing. I next investigated the genomic organization of SLTS genes. 16/20 copies occurred in clusters of 2–5 copies ranging from 1–21 kb long (gray bars at left of fig. 1*B*). Interestingly, in several clusters the putative spliced leader sequences were found on either genomic strand. Scrutiny of transcriptomic sequence from the related species *Hanusia phi* revealed many transcripts beginning with the same mature SL sequence (fig. 1*C*). Thus, *G. theta* and *H. phi* utilize spliced leader splicing with a 17-nt mature spliced leader sequence.

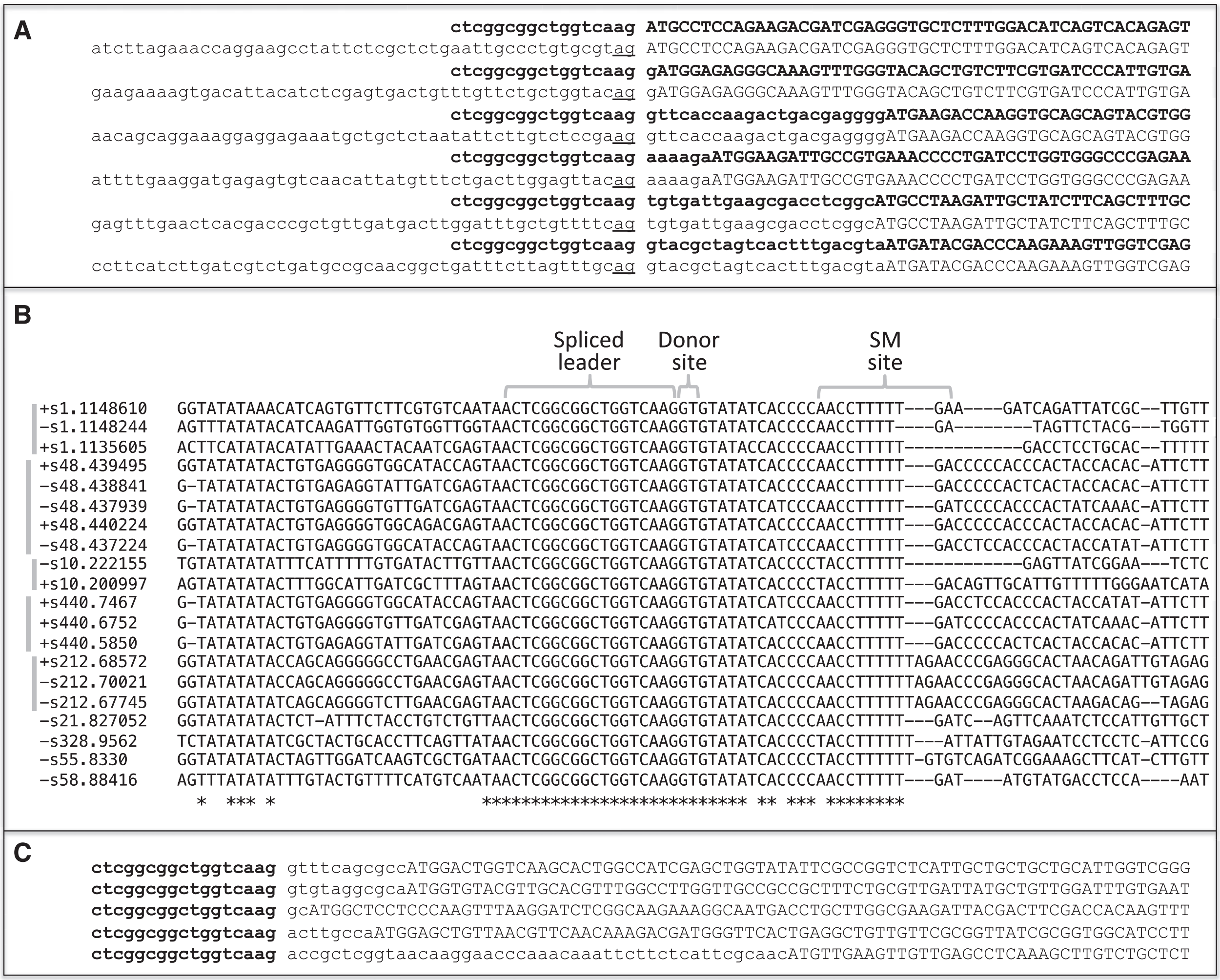
A
ctcggcggctggtcaag ATGCCTCCAGAAGACGATCGAGGGTGCTCTTTGGACATCAGTCACAGAGT
atcttagaaaccaggaagcctattctcgctctgaattgccctgtgcgtag ATGCCTCCAGAAGACGATCGAGGGTGCTCTTTGGACATCAGTCACAGAGT
ctcggcggctggtcaag gATGGAGAGGGCAAAGTTTGGGTACAGCTGTCTTCGTGATCCCATTGTGA
gaagaaaagtgacattacatctcgagtgactgtttgttctgctggtacag gATGGAGAGGGCAAAGTTTGGGTACAGCTGTCTTCGTGATCCCATTGTGA
ctcggcggctggtcaag gttcaccaagactgacgaggggATGAAGACCAAGGTGCAGCAGTACGTGG
aacagcaggaaaggaggagaaatgctgctctaatattcttgtctccgaag gttcaccaagactgacgaggggATGAAGACCAAGGTGCAGCAGTACGTGG
ctcggcggctggtcaag aaaagaATGGAAGATTGCCGTGAAACCCCTGATCCTGGTGGGCCCGAGAA
attttgaaggatgagagtgtcaacattatgtttctgacttggagttacag aaaagaATGGAAGATTGCCGTGAAACCCCTGATCCTGGTGGGCCCGAGAA
ctcggcggctggtcaag tgtgattgaagcgacctcggcATGCCTAAGATTGCTATCTTCAGCTTTGC
gagtttgaactcacgacccgctgttgatgacttggatttgctgttttcag tgtgattgaagcgacctcggcATGCCTAAGATTGCTATCTTCAGCTTTGC
ctcggcggctggtcaag gtacgctagtcactttgacgtaATGATACGACCCAAGAAAGTTGGTCGAG
ccttcatcttgatcgtctgatgccgcaacggctgatttcttagtttgcag gtacgctagtcactttgacgtaATGATACGACCCAAGAAAGTTGGTCGAG
B
Spliced leader
Donor site
SM site
+s1.1148610 GGTATATAAACATCAGTGTTCTTCGTGTCAATAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTT---GAA----GATCAGATTATCGC--TTGTT
-s1.1148244 AGTTTATATACATCAAGATTGGTGTGGTTGGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTT----GA--------TAGTTCTACG---TGGTT
+s1.1135605 ACTTCATATACATATTGAAACTACAATCGAGTAACTCGGCGGCTGGTCAAGGTGTATACCACCCCAACCTTTTT------------GACCTCCTGCAC--TTTTT
+s48.439495 GGTATATATACTGTGAGGGGTGGCATACCAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTT---GACCCCCACCCACTACCACAC-ATTCTT
-s48.438841 G-TATATATACTGTGAGAGGTATTGATCGAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTT---GACCCCCACTCACTACCACAC-ATTCTT
-s48.437939 G-TATATATACTGTGAGGGGTGTTGATCGAGTAACTCGGCGGCTGGTCAAGGTGTATATCATCCCAACCTTTTT---GATCCCCACCCACTATCAAAC-ATTCTT
+s48.440224 GGTATATATACTGTGAGGGGTGGCAGACGAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTT---GACCCCCACCCACTACCACAC-ATTCTT
-s48.437224 G-TATATATACTGTGAGGGGTGGCATACCAGTAACTCGGCGGCTGGTCAAGGTGTATATCATCCCAACCTTTTT---GACCTCCACCCACTACCATAT-ATTCTT
-s10.222155 TGTATATATATTTCATTTTTGTGATACTTGTTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCTACCTTTTT------------GAGTTATCGGAA---TCTC
+s10.200997 AGTATATATACTTTGGCATTGATCGCTTTAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCTACCTTTTT---GACAGTTGCATTGTTTTTGGGAATCATA
+s440.7467 G-TATATATACTGTGAGGGGTGGCATACCAGTAACTCGGCGGCTGGTCAAGGTGTATATCATCCCAACCTTTTT---GACCTCCACCCACTACCATAT-ATTCTT
+s440.6752 G-TATATATACTGTGAGGGGTGTTGATCGAGTAACTCGGCGGCTGGTCAAGGTGTATATCATCCCAACCTTTTT---GATCCCCACCCACTATCAAAC-ATTCTT
+s440.5850 G-TATATATACTGTGAGAGGTATTGATCGAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTT---GACCCCCACTCACTACCACAC-ATTCTT
+s212.68572 GGTATATATACCAGCAGGGGGCCTGAACGAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTTTAGAACCCGAGGGCACTAACAGATTGTAGAG
-s212.70021 GGTATATATACCAGCAGGGGGCCTGAACGAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTTTAGAACCCGAGGGCACTAACAGATTGTAGAG
-s212.67745 GGTATATATATCAGCAGGGGTCTTGAACGAGTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTTTAGAACCCGAGGGCACTAAGACAG--TAGAG
-s21.827052 GGTATATATACTCT-ATTTCTACCTGTCTGTTAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTT---GATC--AGTTCAAATCTCCATTGTTGCT
-s328.9562 TCTATATATATCGCTACTGCACCTTCAGTTATAACTCGGCGGCTGGTCAAGGTGTATATCACCCCTACCTTTTTT---ATTATTGTAGAATCCTCCTC-ATTCCG
-s55.8330 GGTATATATACTAGTTGGATCAAGTCGCTGATAACTCGGCGGCTGGTCAAGGTGTATATCACCCCTACCTTTTTT-GTGTCAGATCGGAAAGCTTCAT-CTTGTT
-s58.88416 AGTTTATATATTTGTACTGTTTTCATGTCAATAACTCGGCGGCTGGTCAAGGTGTATATCACCCCAACCTTTTT---GAT----ATGTATGACCTCCA----AAT
* *** * ************************** ** *** ********
C
ctcggcggctggtcaag gtttcagcgccATGGACTGGTCAAGCACTGGCCATCGAGCTGGTATATTCGCCGGTCTCATTGCTGCTGCTGCATTGGTCGGG
ctcggcggctggtcaag gtgtaggcgcaATGGTGTACGTTGCACGTTTGGCCTTGGTTGCCGCCGCTTTCTGCGTTGATTATGCTGTTGGATTTGTGAAT
ctcggcggctggtcaag gcATGGCTCCTCCCAAGTTTAAGGATCTCGGCAAGAAAGGCAATGACCTGCTTGGCGAAGATTACGACTTCGACCACAAGTTT
ctcggcggctggtcaag acttgccaATGGAGCTGTTAACGTTCAACAAAGACGATGGGTTCACTGAGGCTGTTGTTCGCGGTTATCGCGGTGGCATCCTT
ctcggcggctggtcaag accgctcggtaacaaggaacccaaacaaattcttctcattcgcaacATGTTGAAGTTGTTGAGCCTCAAAGCTTGTCTGCTCT

Fig. 1.—Spliced leader trans-splicing in cryptomonads. (*A*) Observed transcriptomic (bold) and corresponding genomic (nonbold) sequences for several *Guillardia theta* genes, showing acceptor sites in the genome (underline) and relationship to predicted coding sequences (capital letters). (*B*) Alignment of 20 genomic copies of gene encoding spliced leader sequence, showing the mature splice leader sequence, donor site, and predicted SM binding site. Locations in genomic sequence are given (s1 indicates "scaffold_1"), with gray bars highlighting genomic clusters of SL genes on both strands (±). (*C*) Observed transcriptomic sequences in *Hanusia phi*, showing SL sequences (bold) and relationship to predicted coding sequences (capital letters).

SLTS Patterns in Cryptomonads

Among 429 observed transcripts that include an extended ORF for which a single SL splice site was observed (i.e., no alternative SLTS; see below), the distance from the 3′ end of the SL to the putative translation start ATG—that is, the putative UTR—ranged from zero nucleotides (the SL exactly precedes the ATG) to several hundred (fig. 2*A*). There is a clear bias towards short UTRs, with 19% of transcripts having UTRs 0–1 nts, and 36.6% ≤10 nts (fig. 2*A*). Results for *Hanusia phi* were similar (fig. 2*A*). Scrutiny of the genomic sequences of the protein-coding sequence revealed a classic acceptor motif, with a C/TAG splice site preceded by a polypyrimidine tract (fig. 2*B*).

Systematic Identification of SLTS Splice Sites

To better survey trans-splicing in *G. theta* I identified individual RNA-seq reads containing full or partial SL sequences and mapped these to the genome (see Materials and Methods). The SL sequence was found in 868,236 reads in a sample of 183 million reads (Curtis et al. 2012), of which 565,556 mapped uniquely to the genome. About 97% of all mapped sites exhibited the expected AG acceptor site, yielding a total of 7,101 putative AG acceptor splice sites. These data confirmed the picture above, with UTRs ranging up to several hundred nucleotides with a strong peak at UTR lengths of 0 nts (fig. 2*A*). About 20.9% of genes exhibited trans-splicing, an estimate that is likely to be an underestimate due to

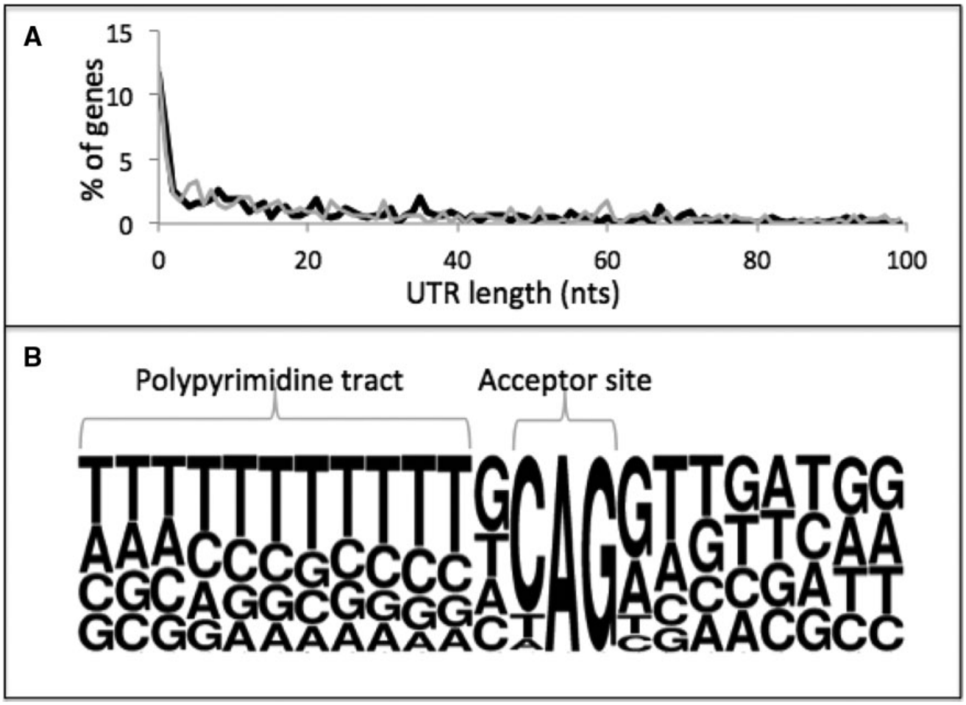

FIG. 2.—Characteristics of spliced leader trans-splicing splice sites in *Guillardia theta* and *Hanusia phi.* (*A*) Distance from donor site to predicted translation start site for trans-spliced genes, showing a wide range of lengths with a strong peak near zero nucleotides (see also fig. 1*A*). Black and gray lines show data for *G. theta* and *H. phi*, respectively. (*B*) Weblogo for observed SLTS acceptor sites in *G. theta*, showing classic features of spliceosomal acceptor sites.

the small amount of RNA-seq data available, possible problems with the annotion of 5′ ends of genes, and the possibility of conditional SLTS of some genes.

Alternative SLTS Sites are Frequent in *G. theta* but Frequently Used at Low Levels and without Impact on Coding Sequences

I next explored alternative SLTS, in which different transcript isoforms of a gene are produced by alternative addition of an SL at different sites (thus each transcript isoform has an SL added at one of two or more sites; Horiuchi and Aigaki 2006; Nilsson et al. 2010). To identify all SLTS sites within each gene, I mapped SL-containing RNA-seq reads to the genome. This revealed that many genes showed multiple splice sites, some examples of which are given in figure 3*A*. In total, 27.3% of all genes exhibiting SLTS had multiple SLTS sites, including 6.9% in which at 10 RNA-seq reads supported each of two or more splice sites (fig. 3*B*). To further characterize alternative SLTS, for each gene exhibiting alternative SLTS I compared the position of the major splice site (that is, the most commonly used splice site for a given gene) with those of the minor splice site(s) (defined as a less frequently used site, but one supported by at least 10 reads). This comparison showed a clear preference for minor SLTS sites to be located near major SLTS sites, in particular a separation of exactly three nucleotides (so-called NAGNAG acceptor sites; Bradley et al. 2012). On the other hand, many other minor splice sites were found dozens or hundreds of nucleotides away from the major site (fig. 3*C*). Notably, though, for the vast majority of genes, a single SLTS site dominated: among genes with at least 10 reads, >90% of reads mapped to the same site for 89.3% of genes (fig. 3*D*).

Interestingly, and contrary to results from other species (Nilsson et al. 2010; Rettig et al. 2012), very few alternative SLTS sites were observed within predicted coding sequences (181 genes out of 3601 genes with a SLTS site supported by at least 10 reads). Among these genes only 23/181 also showed usage of a trans-splicing site upstream of the annotated translation start site, suggesting that these cases reflect misannotation of a single translation start site, downstream of the SLTS splice site. Thus, previous hypotheses that alternative SLTS may act to significantly increase proteome diversity by producing transcripts encoding alternative protein isoforms do not appear to be applicable in cryptomonads (Nilsson et al. 2010).

No Evidence for Widespread Operon Structure in *G. theta*

Other species with known spliced leader transcription exhibit a high degree of polycistronic transcripts, with polycistronic pre-mRNAs being resolved into individual mature mRNAs each

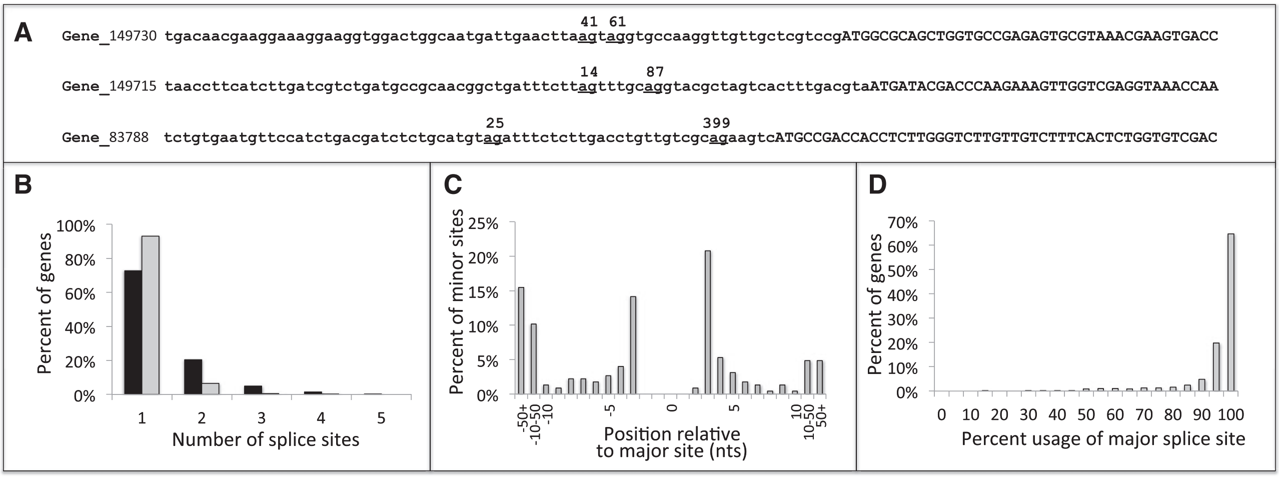

Fig. 3.—Alternative SLTS splicing in *G. theta*. (*A*) Three examples of alternative SLTS in *G. theta*, showing alternative SLTS acceptor sites (underlines). Numbers of reads supporting each splice site are given above the splice site. (*B*) Number of observed splice sites per gene for genes found to be undergoing SLTS, including either all sites (black) or only splice sites supported by at least 10 reads (gray). (*C*) Position of minor trans-splicing sites supported by at least ten reads, relative to major trans-splicing site. Peaks at -3 and +3 represent NAGNAG sites where the major splice site is the second AG and the first AG, respectively. (*D*) Usage of splice sites across all SLTS genes, showing that most genes have a single major splice site accounting for ≥90% of observed trans-spliced reads.

carrying a single protein-coding sequences by trans-splicing (Jäger et al. 2007; Blumenthal et al. 2015). Two signatures of polycistronic transcription are clustering in the genome of trans-spliced genes, and a preference of neighboring trans-spliced genes to reside on the same DNA strand. I found that trans-spliced genes were more likely to be immediate neighbors in the genome than expected at random ($P \leq 0.001$), although the effect was moderate (an excess of 20.7% relative to random expectation, with 645 neighboring pairs of trans-spliced genes, compared to 534 expected at random). Interestingly, among the 645 trans-spliced neighboring pairs, there was no excess of genes transcribed from the same strand ($P > 0.5$), suggesting that the moderate level of clustering of genes reflects processes other than polycistronic transcription.

Concluding Remarks

Spliced leader trans-splicing is known to have evolved many times in different animal lineages, but is only known in a few protist lineages. Here, I report spliced leader trans-splicing in another unicellular protist lineage, cryptomonads. That cryptomonads exhibit particularly complex cell compartmentalization could suggest that spliced leader trans-splicing evolves in the context of greater organismal and/or transcriptomic complexity. However, caution is in order for two reasons. First, the current effort and other efforts to discover trans-splicing have not been systematic taxonomically, thus known instances of SLTS may reflect biased sampling as much as differences in biology. Second, the apparent dearth of commonly used alternative SLTS sites in these data suggests that the potential of SLTS for transcriptome and proteome diversification is not maximized in these species, thus it is unclear to what extent SLTS imparts distinct functions in cryptomonads as opposed to simply representing an alternative mode of gene expression. The reported SLTS in cryptomonads also differs from other characterized lineages in the lack of apparent polycistronic transcription. This breaks what had previously been seen as a tight coupling between SLTS and polycistronic transcription. Possibly, necessary coordination between 3′ processing of one mRNA with 5′ processing of the downstream coding sequence (Jäger et al. 2007) has yet to evolve in cryptomonads.

The results reported here extend the evolution of SLTS to yet another major group of eukaryotes, deepening the mystery of the recurrent evolution of SLTS. With the availability of genomic and transcriptomic resources for wide varieties of eukaryotic lineages, the possibility now exists for a truly general survey of SLTS across organisms. Such a study would allow for testing of hypotheses of the correlates of SLTS and yield insights into the function of this mysterious mechanism.

Materials and Methods

The assembled transcriptomes for *G. theta* and *H. phi* were downloaded from the MMETSP website (http://marinemicroeukaryotes.org/). Version 1 of the genome and genome annotation of *G. theta* was downloaded from the Joint Genome Institute (http://genome.jgi.doe.gov/Guith1/Guith1.home.html). *Guillardia theta* RNA-seq Illumina reads were downloaded from the Short Read Archive at NCBI (http://www.ncbi.nlm.nih.gov/sra, run SRR747855). RNA-seq reads beginning with all or part of the SL sequence (at least 8 nucleotides) were identified, the putative SL sequence removed, and the remainder of the read mapped against the genome, with sequences mapping multiple times being discarded. Reads were then mapped to a database containing annotated protein-coding regions joined to the 1000 nucleotides upstream of the start site. Presence of a genomic AG dinucleotide at the boundary was assessed for each putative splice site, leading to 7101 total observed splice sites. The numbers of times that each splice site was observed were tallied. Genomic copies of the SL gene sequence were identified by performing a local BLASTN search of the mature SL sequence against the genome. A weblogo of SLTS acceptor sites was generated at http://weblogo.berkeley.edu/logo.cgi with default parameters. Tests for clustering of trans-spliced genes and for tendency of neighboring trans-spliced genes to fall on the same genomic strand were performed by randomizing gene order within each scaffold in the genome. Novel analyses were performed with custom Perl scripts.

Literature Cited

Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR. 2013. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. Front Genet. 4:199.

Blumenthal T, Davis P, Garrido-Lecca A. 2015. Operon and non-operon gene clusters in the *C. elegans* genome. WormBook 28:1–20.

Bradley RK, Merkin J, Lambert NJ, Burge CB. 2012. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol. 10:e1001229.

Curtis BA, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59–65.

Douris V, Telford MJ, Averof M. 2010. Evidence for multiple independent origins of trans-splicing in Metazoa. Mol Biol Evol. 27:684–693.

Ganot P, Kallesøe T, Reinhardt R, Chourrout D, Thompson EM. 2004. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. Mol Cell Biol. 24:7795–7805.

Günzl A, et al. 2003. RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. Eukaryotic Cell 2:542–551.

Horiuchi T, Aigaki T. 2006. Alternative trans-splicing: a novel mode of pre-mRNA processing. Biol Cell 98(2):135–140.

Jäger AV, De Gaudenzi JG, Cassola A, D'Orso I, Frasch AC. 2007. mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes. Proc Natl Acad Sci U S A. 104:2035–2042.

Johnson PJ, Kooter JM, Borst P. 1987. Inactivation of transcription by UV irradiation of *Trypanosoma brucei* provides evidence for a multicistronic transcription unit including a VSG gene. Cell 51:273–281.

Keeling PJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 12:e1001889.

Kozak M. 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234:187–208.

Krause M, Hirsh D. 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. Cell 49:753–761.

Kuersten S, Lea K, MacMorris M, Spieth J, Blumenthal T. 1997. Relationship between 3′ end formation and SL2-specific trans-splicing in polycistronic *Caenorhabditis elegans* pre-mRNA processing. RNA 3:269–278.

Lasda EL, Blumenthal T. 2011. Trans-splicing. Wiley Interdiscip Rev RNA 2:417–434.

LeBowitz JH, Smith HQ, Rusche L, Beverley SM. 1993. Coupling of poly(A) site selection and trans-splicing in Leishmania. Genes Dev. 7:996–1007.

Lee MG, Van der Ploeg LH. 1997. Transcription of protein-coding genes in trypanosomes by RNA polymerase I. Annu Rev Microbiol. 51:463–489.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A. 100:189–192.

Maeso I, Roy SW, Irimia M. 2012. Widespread recurrent evolution of genomic features. Genome Biol Evol. 4(4):486–500.

Moore CE, Archibald JM. 2009. Nucleomorph genomes. Annu Rev Genet. 43:251–264.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457–463.

Nilsson D, et al. 2010. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. PLoS Pathog. 6:e1001037.

Nowack EC, et al. 2016. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. Proc Natl Acad Sci U S A. 113:12214–12219.

Peng Z, et al. 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol. 30:253–260.

Rettig J, Wang Y, Schneider A, Ochsenreiter T. 2012. Dual targeting of isoleucyl-tRNA synthetase in *Trypanosoma brucei* is mediated through alternative trans-splicing. Nucleic Acids Res. 40:1299–1306.

Saito TL, et al. 2013. The transcription start site landscape of *C. elegans*. Genome Res. 23:1348–1361.

Shepard PJ, et al. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 17:761–772.

Sutton RE, Boothroyd JC. 1986. Evidence for trans splicing in trypanosomes. Cell 47:527–535.

Yang F, et al. 2015. Spliced leader RNA trans-splicing discovered in copepods. Sci Rep. 5:17411.

Zhang H, et al. 2007. Spliced leader RNA trans-splicing in dinoflagellates. Proc Natl Acad Sci U S A. 104:4618–4623.

Zorio DA, Cheng NN, Blumenthal T, Spieth J. 1994. Operons as a common form of chromosomal organization in *C. elegans*. Nature 372:270–272.

Associate editor: Kenneth Wolfe