Evolution of pre-mRNA spliced-leader (SL) *trans*-splicing in the deuterostomes and its role in monocistronic gene expression

Liane Levasseur
Biology Department
McGill University
December 2012

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science, Biology.

© Liane Levasseur 2012

ABSTRACT

Spliced-leader (SL) *trans*-splicing is a splicesomal process by which the 5'-ends of diverse mRNAs are replaced by a common sequence originating from a specialized SL donor RNA. The patchy phylogenetic distribution of SL *trans*-splicing, including its recent discovery in the deuterostomes in the tunicate *Ciona intestinalis* presents two equally possible evolutionary histories. Either it is an ancestral eukaryotic trait that has been lost in multiple lineages, or, it was invented *de novo* in various lineages including the tunicates. SL *trans*-splicing has a known function in the resolution of polycistronic operons. However, for monocistronic transcripts there are many proposed but not substantiated functions, including the 5'-untranslated region (5'-UTR) sanitization hypothesis. I investigated the phylogenetic distribution of SL *trans*-splicing in the deuterostomes and its possible function as a 5'-UTR sanitization mechanism in the monocistronic troponin I gene (CiTnI) in *Ciona*.

I produced oligo-capped 5'-RACE cDNA libraries to survey mRNA 5'-end sequences for common candidate SL sequences in the cephalochordate *Branchiostoma*, the hemichordate *Saccoglossus* and the echinoderm *Strongylocentrotus*. Following steps taken to eliminate short primer-related sequences created during the 5' oligo-capping reaction, and apparently hidden in the form of heteroduplex complexes, the cDNA libraries were subjected to 454 ultra high-throughput sequencing. Several indicators, including the presence of 5'-TOP sequences at the 5'-termini of mRNAs encoding ribosomal proteins and translation factors showed that the 5'-RACE libraries effectively represented the extreme 5'-termini of mRNAs. No common 5'-ends that could represent SL sequences were observed in any of the species examined. This result strongly supports the independent evolution of SL *trans*-splicing in the tunicates.

To further examine the postulated functional role of SL *trans*-splicing in removing deleterious sequences from the 5'-UTR, I made and used an internal transfection control construct to validate the observation, previously made in our lab, that an experimental CiTnI reporter construct with a long retained-outron 5'-UTR showed low reporter gene activity when electroporated in *Ciona* embryos. I then created a new CiTnI construct to directly assess the possible presence of a specific deleterious element in the long retained-outron 5'-UTR. My results clearly indicated that there is no specific deleterious element, but that it is the length of the retained-outron 5'-UTR, *per se*, that leads to low TnI reporter construct expression.

ABSTRAIT

Le “spliced-leader” (SL) *trans*-épissage est un processus dépendant du spliceseome dont l’éxtremité 5’ de divers ARNms sont remplacés par une séquence commune dérivant d’un SL ARN donneur spécialisé. À cause de sa répartition phylogénétique compliquée et sa découverte récente dans le tunicier *Ciona intestinalis*, un deutérostomiens, il existe deux possibilités d’évolution plausible pour le SL *trans*-épissage. Le SL *trans*-épissage est soit; un caractère ancestral qui s’est perdu plusieurs fois dans différentes lignées où il s’est inventé *de novo* dans différentes lignées, y compris les tuniciers. Une des fonctions connues du SL *trans*-épissage est la résolution de transcrits polycistroniques, cependant sa fonction concernant les transcrits monocistroniques est moins évidente. Il existe plusieurs hypothèses en ce qui concerne la fonction du SL *trans*-épissage de gènes monocistroniques, une d’entre-elles est l’assainissement des régions 5’ non-traduit (5’-UTR). Mes travaux de recherche ont portés sur la répartition phylogénétique de SL *trans*-épissage dans les deutérostomiens et sa fonction en tant d’assainissement de la région 5’ non-traduite pour le gène troponine I (CiTnI) de *Ciona*.

J’ai créé des banques d’ADNc 5’-RACE coiffés d’oligonucleotide pour bien étudier les séquences à l’extrémité 5’ de différents ARNms et ainsi découvrir un motif SL commun dans le *Branchiostoma* céphalochordé, le *Saccoglossus* hemichordate et le *Strongylocentrotus* échinodermes. Après avoir éliminé les courtes séquences reliées à l’amorce qui se sont crée au cours de la réaction de coiffe des ARNs et qui étaient apparemment cachées sous forme de complexes hétéroduplexes, les banques d'ADNc ont été soumises au séquençage 454 ultra à haut débit. La présence des séquences 5’-TOP sur les ARNms des proteines ribosomique et gènes de facteurs de traduction était un des indicateurs confirmant que les banques comprenaient des extrémités 5’ de ARNm de

qualité. Des extrémités 5’ en communs, qui pourraient représenter des séquences SL, n’ont pas été observées dans les organismes examinés. Ce résultat appuie fortement l'évolution indépendante de SL *trans*-épissage dans les tuniciers.

Pour examiner davantage le rôle d'élimination des séquences néfastes de la 5'-UTR postulé pour le SL *trans*-épissage, j’ai produit et utilisé un contrôle interne de transfection. Ce contrôle a été construit afin de valider des résultats obtenus il y a quelques années dans notre laboratoire, qu'un rapporteur expérimental de CiTnI qui contient un outron-5'-UTR non-épissé a une faible activité du gène rapporteur lorsqu’électroporé dans des embryons de *Ciona*. Ensuite, j’ai créé un nouveau rapporteur CiTnI pour évaluer la présence possible d’un élément néfaste dans la séquence du 5’-UTR. Mes résultats indiquent qu'il n'y a aucun élément néfaste spécifique, mais que la longueur du outron-5'-UTR retenu, en soi, mène à la faible expression du rapporteur TnI.

ACKNOWLEDGMENTS

First and foremost I would like to thank my supervisor, Dr. Ken Hastings for allowing me the opportunity to work in his lab and whose patience, invaluable guidance and enthusiasm in this project has been unwavering. I would also like to extend my sincere thanks to all those who were or are currently part of the Hastings' lab, especially our research associate, Dr. Pat Hallauer for her help and advice regarding lab protocols. I would like to thank Dr. Tom Meedel for allowing me to work on *Ciona* in his lab at Rhode Island College (RIC) and his expert assistance in my gene reporter experiments. Additionally, I would also like to thank Drs. Eric Hall and Eric Roberts at RIC for access to their labs and microscopes and their exceptional help in the microscopy of my embryos. I would like to thank my Supervisory Committee; Dr. Ken Dewar and Dr. Richard Roy, for their thoughtful questions and suggestions. I would like to thank the Genome Centre, in particular Dr. Alfredo Staffa, for their time and work on the high-throughput sequencing project. I would like to extend my deepest gratitude to our collaborators for the bioinformatics on the high throughput data; Dr. Ken Dewar and Jessica Wasserscheid. I am grateful to Drs. Nick and Linda Holland and Dr. Ken Hastings for providing *Branchiostoma* RNA, Drs. Andy Cameron and Eric Davidson for providing *Strongylocentrotus* RNA, Drs. Marcos Costa and Marianne Bronner-Fraser for providing *Petromyzon* embryo RNA and Drs. Bob Freeman and Marc Kirschner for providing *Saccoglossus* RNA. I would like to thank everyone from the labs on the 6th floor of the Montreal Neurological Institute for their comradery over the years and for allowing me unlimited access to equipment, reagents and knowledge of such. Finally, to my family and all my friends (including my Neuro colleagues, Bryce, Craig and Celina), I could never have done this without your support, thank you.

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION

OVERVIEW OF THESIS

This thesis is a contribution to research aimed at understanding the biological significance of evolutionary change in fundamental gene expression mechanisms. The specific subject of this work is spliced-leader (SL) *trans*-splicing, a gene expression mechanism found in some, but not all, eukaryotic groups. Arising from discovery of SL *trans*-splicing in the chordates in our laboratory (Vandenberghe, Meedel, & Hastings, 2001), the present research aims to establish the phylogenetic distribution of SL *trans*-splicing within the deuterostomes, the major division of the metazoa that includes the chordates, and also to experimentally elucidate the functional roles of SL *trans*-splicing and its implications for gene structure/expression and for evolutionary genomics.

BACKGROUND

Evolution of gene expression mechanisms

Not only genes, but also fundamental gene expression mechanisms, evolve and this evolution has important implications for genomic and organismal evolution.

The differences between how prokaryotes and eukaryotes process the genetic information in DNA to produce functional proteins demonstrates the importance of evolutionary change in gene expression mechanisms. The primary transcripts of protein-coding genes of eukaryotes can be distinguished from those of prokaryotes based on two main characteristics; a cap structure at the 5'-end and introns.

The cap protects the 5'-end from exoribonuclease degradation and plays a key role in translation (Shuman, 2002). The cap is required for eukaryotic, and not prokaryotic, translation, due to a radical evolutionary change in the mechanism of

ribosome recruitment to mRNA. In prokaryotes the Shine-Dalgarno sequence, an eight nucleotide sequence near the mRNA translation start site that base pairs with the 16S ribosomal RNA drives correct positioning of the mRNA with the translational machinery (Shuman, 2002). In eukaryotes, the initiation of protein synthesis is dependent on the presence of a cap structure at the 5'-end of the mRNA. The cap is recognized by a eukaryotic initiation factor, eIF4E, a component of the eIF4F complex which interacts with the preinitiation complex (PIC), which includes the Met-tRNAi loaded 40S ribosome subunit along with other elongation factors (Sonenberg & Hinnebusch, 2009). Once positioned at the 5'-end of the mRNA, the PIC is poised to begin scanning for the AUG start codon and once found, the scanning arrests and the 60S ribosomal subunit is recruited. This newly formed 80S initiation complex can now continue and complete the protein synthesis (Sonenberg & Hinnebusch, 2009).

This evolutionary change in translation initiation had consequences for genome organization and evolution. Prokaryote genomes have multigene operons transcribed as single RNA molecules, with independent translation of each protein-coding cistron by initiation at internal sites (Kozak, 2005). However, with the evolution of cap-dependent translation in eukaryotes such internal initiation is no longer possible so that eukaryotic genes in general are monocistronically transcribed, and are not organized in operons (Kozak, 2005). Nevertheless, further evolution within the eukaryotes has permitted, in special cases, operon organization (Blumenthal, 2004). A less generally used cap-independent mechanism of translation, internal ribosome entry sites, has also evolved in eukaryotes and may be capable of internal translational initiation in a small number of bicistronic operons (Blumenthal, 2004). More commonly observed are eukaryotes that carry out SL *trans*-splicing (see below) to cleave polycistronic operon transcripts into separate monocistronic mRNAs which acquire a 5'-cap during *trans*-splicing and can be

translated by the general cap-dependent mechanism (Plank & Kieft, 2012; Blumenthal, 2005)

Non-coding, intronic DNA separating the protein-coding exons, the other main feature of eukaryotic transcripts, necessitates a mechanism, the spliceosome, to remove these introns. It is generally believed that the spliceosome must have evolved before intron proliferation, or that introns and the spliceosome must have coevolved, as initial introns would have been deleterious to protein synthesis (Poole, Jeffares, & Penny, 1998; Lynch & Kewalramani, 2003). An additional relevant mechanism, nonsense-mediated decay, a process by which mRNAs with premature stop codons are degraded before they can be translated into potentially deleterious truncated proteins, is thought to have evolved to protect cells from a significant failure rate of intron removal (Chang, Imam, & Wilkinson, 2007; Lynch & Kewalramani, 2003).

The presence of splicing in eukaryotes allowed for the evolution of alternative splicing, a vital gene regulatory mechanism (Mattick, 2004; Lynch & Kewalramani, 2003). Alternative splicing allows a single pre-mRNA to generate multiple different mRNA by splicing-in different combinations of exons. This capacity likely evolved from weak splice site signals allowing those exons to occasionally be skipped, and/or evolution of splicing regulatory factors, such as RNA-binding SR proteins, so that, when bound close to a constitutively spliced exon increased the probability of its skipping (Ast, 2004; Boue, Letunic, & Bork, 2003).

Alternative splicing and nonsense-mediated decay are clear examples of gene expression mechanisms whose evolution was central to the evolution of eukaryotes. Indeed, alternative splicing and other eukaryotic RNA editing features were likely a major contributing factor facilitating the evolution of multicellular eukaryotes (Mattick, 2004).

Spliced-leader *trans*-splicing

The experimental focus of this thesis is SL *trans*-splicing, a specialized form of RNA splicing. SL *trans*-splicing is a two-step spliceosomal process similar to *cis*-splicing used for intron removal during mRNA maturation (Figure 1). The major difference between *cis*-splicing and *trans*-splicing is that *cis*-splicing consists of splicing together two exons present on the same transcript, while *trans*-splicing is a process by which two independently transcribed RNA species, the spliced-leader-donor SL RNA and the target pre-mRNA, are spliced together. In SL *trans*-splicing, the SL RNA has a splice donor site but no downstream acceptor site and the pre-mRNA has an unpaired acceptor site as the 5'-most splicing signal (Nilsen, 1993). Both *cis*- and *trans*-splicing processes begin by the 2' hydroxyl of the branchpoint A nucleotide, present 20 – 40 nt upstream of the splice acceptor site, performing a nucleophilic attack on the 5' phosphate of the G nucleotide of the GU splice donor site (Padgett et al., 1986; Nilsen, 1993; Davis, 1996). This produces a branched intermediate, Y-shaped for *trans*-splicing, and in the shape of a lariat for *cis*-splicing. In both cases the G is now linked to the A in a 2'-5' phosphodiester linkage and the SL/upstream exon contains a free 3' hydroxyl group (Nilsen, 1993). The second step in splicing consists of this 3' hydroxyl conducting a nucleophilic attack on the 3' phosphate of the G of the AG acceptor site (Nilsen, 1993). Thus the SL/upstream exon is now ligated to the downstream exon, releasing in the case of SL *trans*-splicing, the original 5'-end segment of the pre-mRNA, termed the outron, or in the case of *cis*-splicing, the intron (Conrad et al., 1991). Because many pre-mRNAs are SL *trans*-splicing targets, the net result of *trans*-splicing is that numerous mRNA species share a common 5'-terminal sequence, the SL.

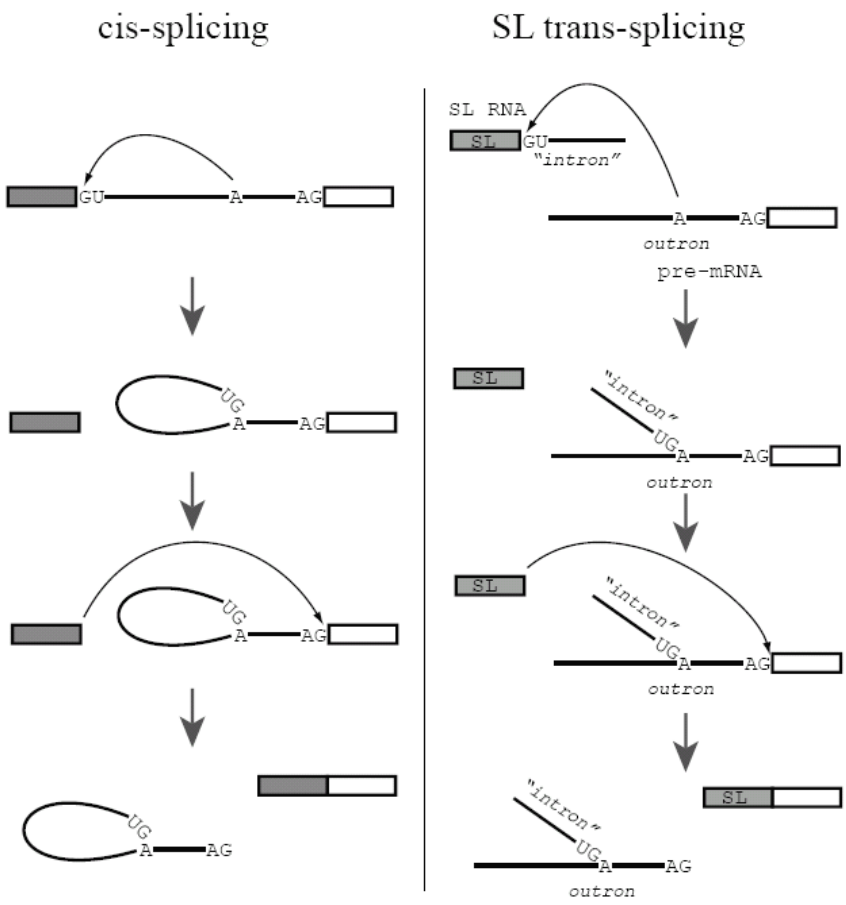


Figure 1: Schematic comparing *cis*-splicing and SL *trans*-splicing. The major steps in the splicing mechanism are shown (see main text for explanation). The end products of *cis*-splicing are a processed, mature mRNA and excised introns as lariat structures. The end products of *trans*-spliced are a SL trans-spliced mRNA and a discarded Y-branched outron. (Figure prepared by K. Hastings, used with permission).

Mechanism of SL *trans*-splicing

SL *trans*-splicing is a spliceosomal process. The spliceosome is composed of multiple uridine-rich, small nuclear ribonucleoproteins (snRNPs) subunits; U1, U2, U5 and U4/U6, and other protein splicing factors (Will & Luhrmann, 2001). Although necessary for *cis*-splicing, the U1 snRNP is not required for *trans*-splicing, but, interestingly, the SL RNA, like small nuclear RNAs (snRNAs), is small, interacts with Sm antigens and contains at its 5'-end a trimethyl guanosine (TMG) cap, and has thus

been postulated to function in the place of U1 for SL *trans*-splicing (Doren & Hirsh, 1988; Hannon, Maroney, & Nilsen, 1991; Thomas, Conrad, & Blumenthal, 1988). However, a more recent study in the nematode *Caenorhabditis* suggests in the "role reversal" model that it is the pre-mRNA that plays the role of the U1 snRNP in *trans*-splicing (Lasda, Allen, & Blumenthal, 2010). In nematodes a U-rich element consisting of a stem loop followed by a UAYYUU motif, is present about 50 nucleotides upstream of the *trans*-splice site in polycistronic operon transcripts (Lasda et al., 2010). Similar to the U1 snRNP motif (GUCCAUUCAUA) role in binding to the 5' splice site of introns in *cis*-splicing, this element is believed to recruit the SL to the pre-mRNA by hybridization with the SL 5' splice site (Lasda et al., 2010).

The different components of the spliceosome perform specific functions in the splicing reaction (Thomas et al., 1988; Nilsen, 1993). The crucial U4/U6 snRNP interacts with both the SL RNA and U2 (Nilsen, 1993). In turn, the U2 snRNP recognizes and binds to the branchpoint sequence of the pre-mRNA, while U2 auxiliary factor, U2AF, reinforces the binding of the spliceosome to the pre-mRNA by recognition of the polypyrimidine tract and AG acceptor site (Nilsen, 1993; Zorio & Blumenthal, 1999; Romfo et al., 2001). Meanwhile U5 interacts with U4/U6 in a tri-snRNP complex and, in *cis*-splicing at least, pairs briefly with sequence just upstream of the acceptor site (Nilsen, 1994; Patel & Bellini, 2008).

Characteristics of the SL RNA

SL RNAs are short, 45-150 nt, RNAs composed of a SL exon at the 5'-end followed by a intron-like sequence that will be removed in the *trans*-splicing process (Hastings, 2005). To date, the shortest SL discovered is that of the tunicate, *Ciona intestinalis*, at 16 nt, and the longest that of the primitive flatworm, *Stylochus zebra* at 51

nt (Vandenberghe et al., 2001; Davis, 1997; Yeats et al., 2010). The length of the 3' intron-like moiety of the SL RNA also varies, the shortest being again that of *Ciona* at 30 nt (Vandenberghe et al., 2001) while the longest is found in the SL RNA of the bdelloid rotifers at 81 nt (Pouchkina-Stantcheva & Tunnacliffe, 2005). SL RNAs from different phyla show little or no overall sequence similarity (Hastings, 2005).

SL RNA genes are transcribed by RNA polymerase II and are typically found in direct tandem repeats ranging in length from 264 bp in *Ciona* (Yeats et al., 2010) to 1.35 kb in *Trypanosoma brucei* (Campbell, Thorton, & Boothroyd, 1984). There appear to be approximately 900 copies of the *Ciona intestinalis* SL RNA gene in tandem repeats (Yeats et al., 2010). Interestingly, in many species, including euglenozoans, nematodes and cnidarians, the SL RNA gene repeats also include the 5S rRNA gene (Drouin & de Sa, 1995; Stover & Steele, 2001). However, this arrangement is likely due to independent, random recombination events, and is not a conserved ancestral trait (Drouin & de Sa, 1995; Brehm et al., 2002; Rajkovic et al., 1990; Zayas, Bold, & Newmark, 2005).

Once transcribed in the nucleus the SL RNA is exported to the cytoplasm where, through an ATP dependent process utilizing the assembly factors of the survival of motor neuron (SMN) complex, it associates with the seven core Sm proteins (Bruzik et al., 1988; Pellizzoni, Yong, & Dreyfuss, 2002). Now, as a snRNP, the SL RNA is able to re-enter the nucleus and interact with the U4/U6 snRNP by base-pairing, an interaction that has been hypothesized to either simply allow for spliceosomal association and thus access to splice acceptor sites or contribute in an unknown way to the *trans*-splicing mechanism (Bruzik et al., 1988; Hannon et al., 1992; Hastings, 2005).

SL RNAs, like other RNA pol II transcripts are initially capped with m7G. Metazoan SL RNA caps examined to date have been shown to be further methylated to

form a TMG cap (Brehm et al., 2002; Davis et al., 1997; Ganot et al., 2004; Lall et al., 2004; Rajkovic et al., 1990; Stover & Steele, 2001; Thomas et al., 1988). TMG caps are also found on splicesomal snRNAs, telomerase RNAs and certain small nucleolar RNAs (Thomas et al., 1988; Seto et al.; 1999, Kiss et al., 2006). Unlike the TMG-capped SL RNA of nematodes, SL RNAs in trypanosomes do not have a TMG cap but rather a cap4 structure, formed by cotranscriptional modification to the first 4 nucleotides downstream of a monomethylated m7G cap (Perry, Waktins, & Agabian, 1987; Mair, Ullu, & Tschudi, 2000). Moreover, immunoprecipitation studies carried out in our laboratory have suggested that the SL RNA of *Ciona* intestinalis does not contain a TMG cap, but is most likely m7G (Yeats, 2009). The role of the TMG cap on some SL RNAs is not clear, neither has its function been elucidated in the U-rich snRNAs, except that it can participate in the cytoplasm-to-nucleus transport of snRNPs following Sm assembly and cap hypermethylation (Liou & Blumenthal, 1990; Matera, Terns, & Terns, 2007).

Functions of SL *trans*-splicing

The functions of SL *trans*-splicing are partly understood, though it is likely that major functionalities remain to be discovered/elucidated (Hastings, 2005). There are presently five distinct known or proposed functions of SL *trans*-splicing (Hastings, 2005).

The first and most generally important known function of SL *trans*-splicing is to drive the resolution of polycistronic operon transcripts into individual capped and translatable monocistronic mRNAs (Blumenthal & Gleason, 2003). Operon resolution, originally discovered in trypanosomes where all transcription is polycistronic, was also later observed in nematodes where two SLs were identified, one (SL1) spliced to monocistronic gene transcripts and a different one (SL2), specialized for polycistron

resolution (Johnson, Kooter & Borst, 1987, Blumenthal, 1995). Polycistronic operons resolved by SL *trans*-splicing are also found in *Ciona*, where the same SL is used for operon resolution and for *trans*-splicing of monocistonic genes (Satou et al., 2006; Satou et al., 2008; Matsumoto et al., 2010).

Another clearly understood though highly specialized function of *trans*-splicing is to provide a 5'-cap to transcripts generated by RNA pol I. This function is unique to the trypanosomes, which are the only organisms known to transcribe some protein-coding RNAs by RNA pol I, which produces uncapped RNAs (Lee & Van der Ploeg, 1997). In most eukaryotes RNA pol I is used uniquely for the transcription of ribosomal RNAs which do not have a 5'-cap. Being transcribed by RNA pol II, the SL RNA is capped, and by *trans*-splicing the splicesomal addition of the SL moiety to the 5'-end of protein-coding RNAs, provides the 5'-cap necessary for subsequent translation and expression.

A third function of SL *trans*-splicing is unique to the platyhelminths, where the last three bases of the SL are a conserved AUG (Cheng et al., 2006). This methionine codon is utilized by a small number of *Schistosoma* mRNAs as the translation start codon. However, as the majority of *trans*-spliced mRNAs do not utilize the SL AUG as the initiation codon, it is highly unlikely that the main function of SL *trans*-splicing is to provide an in-frame AUG (Cheng et al., 2006). Moreover, SL sequences in other phyla do not contain AUG as the last three bases (Cheng et al., 2006).

A fourth proposed function for SL *trans*-splicing is to enhance the translational efficiency of the mRNA, or its stability through the SL sequence specifically or the TMG cap it provides in some species (Liou & Blumenthal, 1990). Enhanced translation could be due to increased mRNA stability, higher affinity between the mRNA and translation factors, or optimization of the 5' untranslated region (5'-UTR) sequence, length or secondary structure (Conrad et al., 1991; Davis, 1996; Lall et al., 2004). There have been

contradicting reports on the effect of *trans*-splicing on mRNA translation. In the nematode *Ascaris lumbricoides*, Maroney et al. (1995) found that the SL sequence, together with the hypermethylated cap present on the SL, collaborate to enhance translation efficiency, most probably by enhancing initiation. However, more recently, Lall et al. (2004) in *Ascaris suum* found that addition of only the SL sequence or only the TMG-cap greatly decreased translation efficiency, while presence of both restored the translation efficiency of the mRNA, although only to approximately the same levels as m7G capped non-SL transcripts. Therefore, it would appear that in *Ascaris suum* at least, SL *trans*-splicing is not utilized to increase translation efficiency. Additionally, by varying the length of a capped 5'-UTR in reporter gene assays, the same group found that optimal in vitro translation of mRNAs in *Ascaris suum* extracts occurred at shorter (25 – 31 nt) 5'-UTR lengths, thus suggesting that addition of the 22 nt SL close to the start codon may serve the function of reducing the lengths of the 5'-UTRs of *trans*-spliced mRNAs by removing the outron portion of the 5'-UTR (Lall et al., 2004).

A fifth proposed function of SL *trans*-splicing is sanitization of the 5'-UTR. The removal of the outron from the pre-mRNA molecule by *trans*-splicing would allow for the evolution of *cis* regulatory elements within the outron which could aid in the initial steps of gene expression, such as transcription, even if the retention of these sequences would have deleterious effects on later steps of gene expression, e.g. mRNA processing, nuclear export, cytoplasmic mRNA stability or translation (Hastings, 2005). Thus a function of SL *trans*-splicing could be to remove these deleterious to the mature mRNA 5'-UTR sequence elements. Alternatively, even in the absence of specific deleterious elements, it might be that it is the overall length of the 5'-UTR that would be deleterious if not shortened by *trans*-splicing. A major issue addressed in this thesis is an experimental assessment of the 5'-UTR sanitization hypothesis (Chapter 4).

Evolution of SL *trans*-splicing

The evolutionary history of SL *trans*-splicing is not clear in part because its known phylogenetic distribution is patchy and the sequence of the SL RNA, though subject to considerable conservation within a phylum, is not conserved across phyla (Davis, 1996). SL *trans*-splicing's similarity to traditional *cis*-splicing and its use of the same spliceosome machinery, make a clear argument for an evolutionary relationship with the *cis*-splicing mechanism, but, whether SL *trans*-splicing was an ancestral eukaryotic trait since lost in multiple lineages or one that appeared *de novo* in the multiple lineages that utilize it is still unknown (Hastings, 2005). Nonetheless, this mechanism is evidently essential for operon resolution and once such *trans*-splicing-dependent operons evolved, it would be very hard to subsequently lose *trans*-splicing in that organismal lineage (Blumenthal, 2005). In the case of monocistronic transcripts SL *trans*-splicing would be similarly difficult to lose once evolved as it removes any pressure to maintain upstream RNA sequence free of in-frame or out of frame start codons or other potentially deleterious elements (Blumenthal, 2005). There are currently no clearly documented cases of either the gain or loss of SL *trans*-splicing during evolution. One of the goals of this thesis research was to assess the possible gain/loss of trans-splicing within the deuterostomes.

Phylogenetic distribution of SL *trans*-splicing

SL *trans*-splicing was originally discovered in the protist euglenozoan, *Trypanosoma brucei* (Campbell et al., 1984), and since then has been found in an additional protist phylum, Alveolata (Zhang et al., 2007) and multiple metazoans; the nematodes (Krause & Hirsh, 1987), platyhelminths (Rajkovic et al., 1990; Brehm et al.,

2002; Zayas et al., 2005), cnidarians (Stover and Steele, 2001; Derelle et al., 2010) rotifers (Pouchkina-Strantcheva & Tunnacliffe, 2005), and tunicates, which are chordates (Vandenberghe et al., 2001; Ganot et al., 2004). Most recently, SL *trans*-splicing has also been discovered in the crustacean arthropods, the ctenophores, and Porifera sponges (Douris, Telford, & Averof, 2010).

The bilaterian metazoa divide into two major branches, the protostomes and the deuterostomes. Most of the metazoan phyla known to carry out SL *trans*-splicing are protostomes. The major deuterostome phyla are the echinoderms, the hemichordates and the chordates, the last consisting of the tunicates, the cephalochordates and the vertebrates (Figure 2). Currently tunicates are the only deuterostomes known to utilize SL *trans*-splicing. Of the tunicate species investigated carefully for SL trans-splicing, including the larvacean *Oikopleura*, colonial ascidian *Botryllus*, and solitary ascidian *Ciona*, all three utilize the trait indicating that SL *trans*-splicing occurs throughout the tunicates (Ganot et al., 2004; Gasparini & Shimeld, 2011).

Based on analysis of 5'-sequences of conventional cDNA expressed sequence tag (ESTs) data from 75 different metazoan species, Douris and colleagues suggested that no additional metazoan phylum beyond those listed above, including deuterostome phyla, did SL *trans*-splicing and that it therefore evolved independently multiple times (Douris et al., 2010). To search for common 5'-ends in the mRNA ESTs from any given species, which could reflect the presence of a spliced-leader sequence, they designed a computer program, Quickmatch. Deuterostome organisms included in their analysis were *Branchiostoma floridae* (a cephalochordate), *Saccoglossus kowalevskii* (a hemichordate) and *Strongylocentrotus purpuratus* (an echinoderm) and no common 5' mRNA sequences that could represent SLs were identified in these species. However, conventional cDNAs/ESTs are likely to be missing at least 8-21 nucleotides at their 5'-

end as a consequence of the mechanism of cDNA second-strand synthesis (D'Alessio & Gerard, 1988). Thus, although a positive finding of SL *trans*-splicing in the conventional cDNA EST analysis of Douris et al. can be conclusive, a negative finding is inconclusive because a short spliced leader sequence could easily escape detection. For example, the 16 nt spliced leader of *Ciona* intestinalis is not detectably present at the 5'-end of conventional ESTs (Satou et al., 2006). Therefore the study of Douris et al. is inconclusive in that, while it likely does eliminate the possibility of long SLs in *Branchiostoma*, *Saccoglossus* and *Strongylocentrotus*, it does not eliminate the possibility of short SLs such as the 16 nt SL of the tunicate *Ciona*. However, direct analysis of mRNA 5' terminal sequences by 5'-RACE methods can identify even the shortest SLs, including the SL of *Ciona* (Satou et al., 2006), the shortest currently known, and provide an approach to definitively assess the possible existence of common 5'-terminal mRNA sequences that could represent *trans*-spliced leaders. An important experimental goal of this thesis research was to employ 5'-RACE analysis to definitively examine the distribution of SL *trans*-splicing in the deuterostomes.

The SL *trans*-splicing status of most deuterostome groups is inconclusive at best. This however, cannot be said of the gnathostomes (the higher vertebrates). High-throughput analysis of sequence data, including 5'-RACE data, from advanced vertebrates (e.g. mouse, human) shows no evidence for SL *trans*-splicing (Nilsen, 2001). Given this lack of evidence for common 5'-ends from the large amount of available mRNA sequence data, it can be concluded that *trans*-splicing is not occurring, or affects only a very limited number of transcripts, in these higher veterbrates (Nilsen, 2001; Hastings, 2005).

In organisms known to utilize SL *trans*-splicing the fraction of genes producing *trans*-spliced mRNA varies from 25% - 100% (Ganot et al., 2004; Campbell et al., 1984).

Currently, the only known group to *trans*-splice all of its pre-mRNA species are the trypanosomes (as discussed above) (Campbell et al., 1984; Parsons et al., 1984). The nematodes *trans*-splice 50% - 90% of their genes, including all operon downstream genes (Zorio et al., 1994; Lasda, Allen, & Blumenthal, 2010). The tunicate *Ciona trans*-splices 50% of its genes (Satou et al., 2006), meanwhile, a different tunicate, the larvacean species, *Oikopleura dioica*, appears to *trans*-splice only 25% of its genes (Ganot et al., 2004).

The deuterostomes

The eukaryotes, one of the three main domains of life, emerged approximately 1100 million years ago (Ma) and are defined by their membrane-bound nucleus, cytoplasmic endomembrane system and cytoskeleton and include all plants, metazoans, fungi and protists (Knoll, 1992; Douzery et al., 2004). According to molecular clock estimates, the last common ancestor of all the metazoans existed 800 Ma, during the Cryogenian Period (Erwin et al., 2011). Also occurring during this period, about 100 million years later, was the emergence of the Bilateria, followed very quickly by its radiation into the protostomes and the deuterostomes (Erwin et al., 2011).

In 2000, Cameron et al. suggested, based on 18S rRNA analysis of free-living versus sessile classes of hemichordates that the ancestral deuterostome was most likely a mobile worm-like organism (Cameron, Garey, & Swalla, 2000). Furthermore, it was probably a filter-feeder with a dorsal nerve cord (Delsuc et al., 2006; Erwin et al., 2011). It is of interest to note that the chordates are a very ancient phylum, emerging early in the Ediacaran Period (635 – 542 Ma), while the evolution of other deuterostome crown group phyla, namely echinoderms and hemichordates, occurred between the end of the Ediacaran (635 – 542 Ma) and the end of the Cambrian (542 – 490 Ma). Since 500 Ma no

new phyla have been established in either the protostomes or deuterostomes (Erwin et al., 2011). Within the chordates the evolution of the vertebral column and thus the emergence of the vertebrates occurred ~525 Ma, while jaws, the main characteristic of the gnathostomes evolved ~380 Ma (Northcutt, 2012).

Our understanding of the evolutionary relationships among the major phyla comprising the deuterostomes has undergone significant revision in recent years. The original view of the deuterostomes, based on morphological criteria, described a progressive evolution to more vertebrate-appearing organisms, with echinoderms at the base of the tree, followed by hemichordates, tunicates, and finally, cephalochordates, then thought to be the most closely related to the vertebrates (Gee, 2006). In 1994, the inclusion of 18S rRNA sequence data with the use of morphological data to determine the phylogenetic distribution of the deuterostomes supported a monophyletic Chordata that is the sister group to a new clade encompassing the echinoderms and hemichordates; the ambulacraria (Tubeville, Schulz, & Raff, 1994). Furthermore, in 2006, Delsuc et al., using genomic sequence data from multiple deuterosomes, presented for the first time a convincing case for the grouping within the chordates of tunicates and vertebrates into a clade named Olfactores (Delsuc et al., 2006). This remodeled the chordates and placed the tunicates as more closely related to vertebrates than are the cephalochordates (Figure 2). This view of the chordate branch of the deuterostome phylogenetic tree remains unchallenged while other aspects have undergone further suggested revision. These latter changes, based on mitochondrial genomes, protein sequence and microRNA complements, include a proposed addition, as a sister group to the Ambulacraria, of a new clade of marine worms Xenacoelomorpha; consisting of a new phylum; Xenoturbellida, and, Aceolmorpha (Bourlat et al., 2006; Philippe et al., 2011).

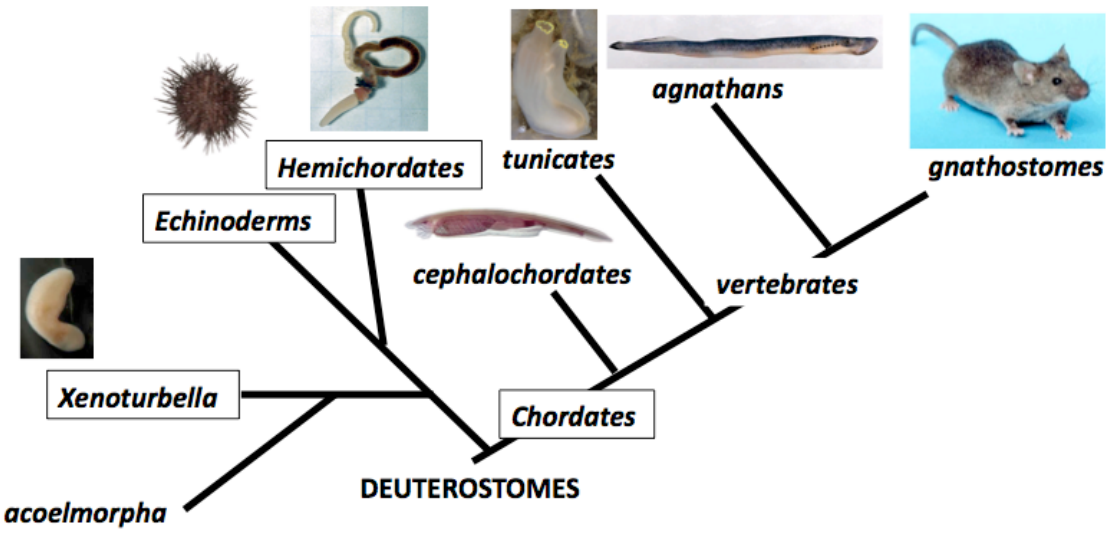


Figure 2: Evolutionary relationships among the deuterostomes. Boxes indicates phyla. Adapted from Cameron, Garey & Swalla, 2000; Bourlat et al., 2006; Delsuc et al., 2006; Phillippe et al., 2011. (Prepared by K. Hastings and L. Levasseur).

Most importantly the current view of the deuterostome tree shows the tunicates as the closest living relative of the vertebrates (Bourlat et al., 2006; Philippe et al., 2011). Given the presence of SL *trans*-splicing in tunicates and its absence in higher vertebrates this sets up a fascinating evolutionary question. Was SL *trans*-splicing present in the tunicates + vertebrates last common ancestor, and subsequently lost in the vertebrate lineage? Or, was SL *trans*-splicing absent in the common ancestor, and invented within the tunicate lineage? To put the question more broadly, what was the history of SL *trans*-splicing in the deuterostomes?

Currently the only deuterostome group known to utilize SL *trans*-splicing is the tunicates, however, if the SL *trans*-splicing distribution within the deuterstomes was known then its status in the tunicate + vertebrate last common ancestor might be elucidated. If, for example, SL *trans*-splicing is found not to occur in any other phylum or chordate subphylum group this would suggest tunicates invented the trait *de novo* within

the deuterosomes. Conversely if SL *trans*-splicing were found in the other deuterostome phyla then this would suggest a loss in the veterbrates.

Ciona intestinalis

Ciona intestinalis is a key species for this thesis research. First, it is the species in which SL *trans*-splicing was first discovered in the deuterostomes (Vandenberghe et al., 2001). Second, it is the species in which I experimentally tested the 5'-UTR sanitization hypothesis.

Ciona intestinalis is an ascidian tunicate. Ascidians are marine animals found all over the world in both shallow and deep waters (Lemaire, 2011). Although adult ascidians are sessile, hermaphroditic, filter-feeders, as larvae they are motile, utilizing their tail muscles and notochord to swim (Satoh, 2003). The larval body plan is very simple, with only a trunk and tail (Satoh & Jeffery, 1995). The *Ciona* tadpole is composed of only ~ 2600 cells and has only a few tissue types, most notably, the epidermis, central nervous system, endoderm and mesenchyme in the trunk and notochord and muscle in the tail (Satoh, 2003). Interestingly, the tail, along with many other larval tissues undergoes apoptosis and resorption during metamorphosis, thus the chordate features of a notochord and dorsal neural tube are not present in the adult tunicate (Satoh & Jeffery, 1995; Lemaire, 2011). The embryogenesis of tunicates is very fast, 18 hours from zygote to larva (Satoh, 2003). Their simple body plan, rapid development and availability of an efficient gene transfer technique (Corbo, Levine, & Zeller, 1997), make ascidians ideal for evolutionary-developmental and gene expression studies (Satoh, 2003).

PROJECT RATIONALE

My project included two main studies. The first was to determine if SL *trans*-splicing is present in any of the main deuterostome groups apart from the tunicates and to thus shed light on the evolutionary question of whether SL *trans*-splicing was lost in the vertebrates or invented in the tunicates. The second study was to further investigate the function of SL *trans*-splicing on monocistronic genes in *Ciona* by probing the 5'-UTR sanitization hypothesis. The rationale for examining both of these facets of SL *trans*-splicing is described below.

Phylogenetic distribution of SL *trans*-splicing

In order to determine if tunicates are the only deuterostomes to utilize *trans*-splicing a methodical protocol to create cDNA including the very first base of the mRNA transcripts and then search for common 5'-end sequences was needed. As discussed above, the use of conventional ESTs as in the approach of Douris et al. (2010) cannot exclude the possibility of shorter SL sequences and would have failed to find *Ciona*'s SL had the sequence not already been known. Therefore, I used a random primed 5'-RACE method, oligo-capping (Maruyama & Sugano, 1994), followed by amplification of the cDNA and high-throughput sequencing to directly analyze mRNA 5' terminal sequences in several deuterostome species.

The four species chosen for my project (*Strongylocentrotus*, Sacoglossus, *Branchiostoma* and *Petromyzon*) represent key deuterostome groups and also have well-developed genome sequencing projects. Xenocoelmorpha was not included based on lack of a sequenced and annotated genome (Philippe et al., 2011). Furthermore the use of these four species from the three traditional phyla of deuterostomes (chordates, hemichordates, and echinoderms) should be sufficient to answer our question. If an SL is

not found in either the echinoderm or hemichordate phyla, or any other chordates, (*Branchiostoma* and *Petromyzon*), then this would suggest tunicates invented SL *trans*-splicing *de novo*. Conversely, if SL *trans*-splicing is found in *Strongylocentrotus purpuratus* (an echinoderm) and *Saccoglossus kowalevskii* (a hemichordate) this would suggest loss in the vertebrates, and the *trans*-splicing status of the primitive vertebrate *Petromyzon* could shed light on when in the vertebrate lineage SL *trans*-splicing was lost. Other combinations of SL distribution would suggest a more complicated evolutionary history.

Function of SL *trans*-splicing

Previous work in this lab discovered, for the first time, the presence of *trans*-splicing in the chordates, more specifically, the tunicate species, *Ciona intestinalis* (Vandenberghe et al., 2001). Our laboratory has characterized and studied in depth the *trans*-spliced *Ciona intestinalis* gene (CiTnI) encoding the muscle contractile regulatory protein troponin I (MacLean, Meedel, & Hastings, 1997; Vandenberghe et al., 2001; Cleto et al., 2003; Khare et al., 2011). In order to better understand the role of SL *trans*-splicing in monocistronic genes, such as CiTnI, and, to specifically test the 5'-UTR sanitization hypothesis, a previous graduate student in our lab attempted to create a non-*trans*-spliced, outron-retaining CiTnI mRNA by deleting the natural *trans*-splice acceptor site (Mortimer, 2007). The unanticipated activation of an upstream cryptic acceptor site did not allow for production of mRNAs retaining the entire outron, however the different experimental constructs generated mRNAs retaining different segments of the outron, and were therefore suitable for assessing the possible presence of deleterious elements (Mortimer, 2007). S. Mortimer's initial results suggested that the mRNA with the longest retained outron segment showed much lower reporter gene expression than those with

shorter retained outron segments (Mortimer, 2007). I wished to confirm her results by including an internal transfection control. Additionally as her results suggested the presence of a deleterious element in the outron of the TnI gene, this was further investigated by generation of a reporter construct giving the same length retained-outron segment, but lacking the putative deleterious sequence.

CHAPTER 2 – METHODS

DNA CLONING AND GEL ELECTROPHORESIS

Standard plasmid cloning protocol

Constructs and plasmids mentioned in this section are listed in Appendix II. All constructs were generated using the same general method (adapted from Invitrogen's Subcloning Efficiency DH5α Competent Cells protocol), with any deviations noted in the description of the specific clone.

All DNA restriction fragments or PCR products to be cloned were isolated using the EZNA Gel Extraction Kit (Omega Bio-Tek) following electrophoresis on a 1% agarose gel.

For the CiTnI and control reporter gene constructs the insert and vector obtained as described for each construct were ligated at a 1:1 molar ratio overnight at 10°C in a 10 µl reaction containing 50 ng total DNA and with 0.05 U of T4 DNA ligase (MBI Fermentas) in T4 DNA ligase buffer, 40mM Tris-HCl, 10 mM $MgCl_2$, 10 mM DTT, 0.5 mM ATP, pH 7.8 (MBI Fermentas).

Oligo-cap 5'-RACE PCR products were cloned using Invitrogen's TOPO TA Cloning Kit. Briefly, 4 µl of DNA (~20 ng) in 2mM Tris, pH 8-8.5 was mixed with a 1 µl 1.2M NaCl, 0.06 M MgCl2 solution and 1 µl (10 ng) pCR4-TOPO linearized vector (Invitrogen) for 20 min at room temperature. The cloning reaction was immediately used for transformations as described below.

Ligated DNA was transformed into competent DH5α cells by a heat shock procedure. Ligation reaction (5 µl) or 2 µl of the TOPO TA cloning reaction, was mixed with 40 µl of sub-cloning efficiency DH5α cells (adapted from Sambrook, Fritsch, & Maniatis, 1989) and incubated on ice for 30 min, followed by a 20 s incubation at 37°-

42°C and a subsequent 2 min incubation on ice. Pre-warmed LB medium (1 ml), or 250 µl room temperature SOC medium was added to the cells followed by incubation for 1 h in a 37°C shaking incubator (Sambrook, Fritsch & Maniatis, 1989). The cells were plated on LB agar+Amp+X-gal+IPTG plates [0.025 µg/ml Ampicillin, 0.025 µg/ml X-gal, 0.023 µg/ml IPTG] and grown overnight at 37°C. The next day the appropriately coloured colonies (blue or white, depending on the ligation) were re-streaked onto new LB agar+Amp+X-gal+IPTG plates. Single colonies from these plates were then used to inoculate 5 ml LB+Amp [25 µg/ml] overnight cultures and plasmid DNA was isolated by a plasmid miniprep procedure (Bio Basic) followed by restriction digests to identify presence and orientation of insert.

For electroporation constructs colonies with correct inserts were then used to make 20% glycerol stocks for long-term storage at -80°C. For each DNA construct used for the electroporation experiments two independent colonies were streaked out from the glycerol stock and grown in culture, from which independent plasmid DNA preparations were made using the Qiagen Maxiprep Kit.

All recombinant DNA clones were sequenced across the ligation junctions by the DNA sequencing facility of the Institute for Research in Immunology and Cancer at Universite de Montreal.

Agarose gel electrophoresis

All DNA, including PCR products, and RNA, was separated by electrophoresis in a 1x TAE [Tris acetate EDTA buffer: 40 mM Tris base (BioShop), 20 mM sodium acetate (BioShop), 2 mM EDTA (BioShop), pH 8.3]. All gels were made with ~1% agarose. Ethidium bromide (Sigma) was included at a final concentration of ~0.5 µg/ml

before pouring and solidification of the gel. Samples were mixed with a 1/10 volume of loading buffer (50% glycerol, 0.1 M EDTA, pH 8, and bromophenol blue) before loading on the gel. After electrophoresis products were visualized by ultraviolet transillumination fluorescence.

CiTnI OUTRON-RETENTION EXPERIMENTS

GFP control construct

A GFP transfection control construct was made, CiTnI(-1.5 kb)eGFP. The construct contained the 1.5 kb *Ciona* troponin I gene promoter region (Cleto, 2002, Khare et al., 2011) driving a promoterless eGFP, in a pBluescript II SK + vector. The DNA was created by first making an intermediate product, pB(CiTnI –1.5), consisting of the CiTnI 1.5 kb promoter region inserted in the multiple cloning site of the pBluescript II SK + vector. This intermediate was made by digesting CiTnI(-1.5 kb)nlacZ [previously named CiTnInuclacZ(-1.5), Cleto, 2002] with KpnI and isolating the 1417 bp CiTnI promoter fragment as described above. The ~1.5 kb fragment was then ligated into KpnI digested pBluescript II SK+ vector and transformed using the standard protocol (see above). Plasmid DNA isolated from white colonies was then digested with XmnI restriction enzyme to assess the orientation of the CiTnI insert. Due to the non-directional ligation used, both possible insert orientations could occur. XmnI cut asymmetrically within the insert and thus permitted discrimination between the two orientations. If the insert were ligated in its forward (correct) orientation the digest would give bands of 2.5 kb and 1.8 kb while in the reverse orientation the bands would be 2.8 kb and 1.5 kb.

The eGFP fragment to be inserted into the intermediate construct, pB(CiTnI –1.5), was gel-recovered as a 2 kb fragment from a XhoI/Eco147I restriction digest of pEGFP-N1 (Clontech). This fragment was ligated into XhoI/EcoRV digested pB(CiTnI –1.5)

DNA and following transformation white colonies were selected. Constructs were verified by restriction digest with XbaI to confirm identity. A digestion pattern of 3.7 kb and 2.6 kb bands would be expected. A colony containing the correct CiTnI(-1.5 kb)eGFP DNA plasmid was grown and processed as described above to obtain two separate sequence-verified DNA preps.

CiTnI(AC)nlacZ outron-retaining construct

The construct CiTnI(AC)nlacZ was engineered to contain a deletion from position -274 to -177 of the CiTnI outron (region B of Figure 26). It was made by overlap extension mutagenesis PCR (Ho et al., 1989), using CiTnI(ABC)nlacZ (Appendix II) as a template. Two PCR products were generated, one upstream of the deletion and crossing the deletion breakpoint using primers 9394 and 9395 and one downstream of it and crossing the deletion breakpoint using primers 9396 and 9397 [See Appendix I for primer sequences]. The upstream and downstream products were separately amplified using iProof DNA Polymerase (Bio-Rad). For each PCR reaction 1 ng of CiTnI(ABC)nlacZ DNA was mixed with 100 µl HF Buffer (Bio-Rad) supplemented with 0.2 mM dNTPs, 10 µM of each primer (9394 and 9395 for the upstream fragment, 9396 and 9397 for the downstream fragment) and 1 U of iProof DNA Polymerase. The thermocycler program was 98°C for 30 s, [98°C for 10 s, 57°C for 20 s, 72°C for 30 s] x25 cycles, 72°C for 5 min. Afterwards the two PCR products were recovered following electrophoresis on a 1% agarose gel and the upstream half (600 bp) and downstream half (900 bp) products were isolated as described above.

Overlapping PCR was done by first combining 50 ng of the upstream and downstream products in 50 µl HF Buffer (Bio-Rad) and 0.2 mM dNTPs and incubating for 3 min at 98°C followed by incubation on ice for 3 min to allow the two

complementary ends (of primers 9395 and 9396) to anneal. After this, 0.5 U of iProof was added to the reaction and extension allowed to occur for 10 min at 72°C. Finally, 10 μM each of the two extreme outer primers, the forward 9394 and reverse 9397, were added to the reaction and PCR completed with the same thermocycler program as above. The 1.5 kb PCR product was then isolated following electrophoresis on a 1% agarose gel and prepared for cloning by sequential digests with a 10-fold excess of SmaI (New England Biolabs) in NEB 4 buffer for 2 h at 25°C. After this digest, a 20-fold excess of PstI (New England Biolabs) was added to the same digest and incubated for 2 additional hours at 37°C. The 639 bp band from the digest, including the deletion, was isolated following electrophoresis on a 1% agarose gel to be used as an "insert". The CiTnI(ABC)nlacZ template DNA was similarly digested and the 6.5 kb band isolated to be used as the "vector".

Standard ligation and heat shock protocols as described above were used. A blue colony containing the correct CiTnI(AC)nlacZ DNA plasmid was grown and processed as described above to obtain two separate sequence-verified DNA preps.

Electroporation of DNA constructs into *Ciona* embryos

I performed the electroporation experiments in the lab of our collaborator, Dr. Thomas Meedel at Rhode Island College, in Providence, Rhode Island, USA. The constructs were transported from Montreal as wet ethanol precipitates at room temperature and dried and resuspended at 4 ug/μl in water upon arrival.

The electroporation procedure was adapted from Corbo, Levine, & Zeller (1997). For each set of electroporations eggs were collected surgically from 5 or more adult *Ciona intestinalis* and rinsed in Millipore-filtered (MPF) sea water. Approximately one

microliter of sperm from at least 2 individuals was added and mixed with the eggs, and fertilization allowed to occur for 5 min in ~50 ml MPF sea water. After fertilization zygotes were spun down with a hand-cranked centrifuge twice to remove excess MPF sea water and sperm. Fertilized eggs were then washed once in 5 ml sodium thioglycolate solution (0.01 g/ml sea water, adjusted to pH 10 with sodium hydroxide). The removal of follicle cells and of the chorion which prevents transformation was achieved by adding another 5 ml aliquot of sodium thioglycolate solution to which had been added 12.5 µl of 100 mg/ml Protease E (Sigma). Dechorionation was allowed to proceed at room temperature (for approximately 2 min) until about 20% of the chorions had been removed, assessed by monitoring small aliquots of the fertilized eggs under a dissecting microscope. Dechorionated zygotes were collected by gentle hand-cranked centrifugation, and then washed twice in 5 ml, followed by twice in 3 ml, of MPF sea water. Dechorionated zygotes were resuspended in MPF sea water and transferred to a 1% agarose (in sea water) coated petri dish until electroporation.

DNA for electroporation was prepared by mixing the appropriate proportions of plasmid DNAs [experimental construct + transfection control construct = 25 µg] with 600 µl of 0.77 mannitol (Sigma). A 200 µl aliquot of dechorinated zygotes in MPF sea water was then added to the DNA/mannitol, mixed and placed in an electroporation cuvette (4 mm gap, VWR). Electroporations were done using a BTX Unit Model EC600 electroporator with a set voltage of 50 V, a capacitance of 850 µF, a resistance of 480 Ω and a time constant of 20 ms. After electroporation zygotes were transferred to a new 1% agarose (in sea water) plate, covered in sea water, and allowed to develop for 12 h at 18°C (mid tailbud stage).

Embryos were sorted under a dissecting microscope. Normal appearing tailbud embryos were collected assessed for GFP expression analysis by fluorescence

microscopy, and then fixed in 4% paraformaldehyde (Sigma), 0.1% Tween 80 (Sigma) in sea water for at least 30 min on ice and washed three times in PBT_{80} [phosphate buffered saline (Sigma) containing 0.1% Tween80 (Sigma)].

GFP detection

Live embryos were placed on glass depression microscope slides and scored positive or negative for GFP expression under an upright epifluorescence microscope (Nikon Eclipse E600, GFP filter: FITC-HYQ). The embryos were then collected from the slides and fixed, as described above, for subsequent co-transfected reporter gene expression analysis.

Some CiTnI(-1.5 kb)eGFP transformed and untransformed embryos were stored, unfixed at 4°C (to delay development) until they could be examined by laser-scanning confocal microscopy, later the same day. GFP expression was assessed from the maximum intensity projection of 40 z-slices using the EGFP Excitation package of the Olympus Fluoview FV1000 inverted microscope.

β-galactosidase staining of tailbud embryos

The PBT_{80} in which the fixed embryos were stored was replaced with β-galactosidase staining solution (0.04% X-gal, 2 mM $MgCl_2$, 0.06 M Na_2HPO_4, 0.04 M NaH_2PO_4, 4 mM potassium ferrocyanide, 4 mM potassium ferricyanide). Embryos were stained in the dark at room temperature for either 10 min or 2.5 h depending on the experiment (see Chapter 4). The X-gal solution was removed, and the embryos washed once in PBT_{80} and stored in PBT_{80} with 10 mM sodium azide at 4°C.

Scoring β-galactosidase activity

The X-gal stained embryos were photographed under a Leica MZ 95 dissecting microscope using a Canon Powershot S40 digital camera and its accompanying Remote Capture software.

To score the gene expression levels of the different CiTnI/nlacZ constructs a 4-tier system was utilized in which the scores reflected the number of positively stained tail muscle cells present in an embryo. Embryos with 6 or more strongly-stained muscle tail cells stained were scored (+++), embryos with 3-5 strongly-stained cells were scored (++), those with only 1-2 faintly stained cells were scored (+) and those with no cells stained were scored as (0) [Cleto et al., 2003; Khare et al., 2011].

CREATION OF OLIGO-CAP 5' RAPID AMPLIFICATION OF CDNA ENDS LIBRARIES AND HIGH-THROUGHPUT SEQUENCING

Preparation of RNA samples

I isolated *Petromyzon* adult RNA from body wall muscle. Frozen pulverized tissue (500 mg) was homogenized in 5 ml of Trizol (Invitrogen). Chloroform, 1 ml, was added to the homogenate followed by a 15 min centrifugation at 10000 g at 4°C. The top phase was removed and incubated in 2.5 ml of isopropyl alcohol for 10 min. The mixture was again centrifuged at 10000 g for 10 min. The supernatant was removed and the pellet was washed in 70% ethanol and spun at 5000 g for 5 min. The supernatant was removed and the pellet allowed to air dry, and was resuspended in 200 µl water and stored at -80°C.

Petromyzon embryo RNA (~50 µg) prepared using RNAqueous (Ambion) kit and sent to Montreal as a wet ethanol pellet was provided by Drs. Marcos Costa and

Marianne Bronner-Fraser of the California Institute of Technology. Upon receipt the RNA was resuspended in 50 µl water and stored at -80°C.

Saccoglossus embryo RNA (~50 µg) was provided by Drs. Bob Freeman and Marc Kirschner from Harvard. Upon receipt it was immediately resuspended in 25 µl water and stored at -80°C.

Strongylocentrotus embryo RNA (~50 µg) was provided by Drs. Andy Cameron and Eric Davidson from the California Institute of Technology. The RNA was sent to Montreal as an ethanol precipitate. Upon receipt it was spun down and resuspended in 50 µl water and stored at -80°C.

Branchiostoma adult and embryo RNA was obtained by Dr. Ken Hastings at Drs. Nick and Linda Holland's summer research lab at the University of South Florida, Tampa. The RNA was extracted as described above for the *Petromyzon* adult muscle sample. The RNAs were transported to Montreal as wet ethanol pellets and upon arrival both were resuspended in 40 µl water and stored at -80°C.

The concentrations of all RNA samples were measured by spectrophotometric absorbance at 260 nm and upon electrophoresis in a 1% agarose gel, all samples showed prominent 18S and 28S ribosomal RNA (rRNA) bands (the latter predominating) and a transfer RNA band.

Oligo-cap ligation to mRNAs

For 5'-RACE analysis, the oligo-cap method developed in 1997 by Suzuki et al. (Invitrogen's GeneRacer Kit, RLM-RACE) was utilized for all 5 species (4 test species plus *Ciona intestinalis* as a control) studied. RNA (5 μg) was first dephosphorylated for 1 h at 50°C with 10 U of calf intestinal phosphatase (CIP) in 10 μl CIP Buffer (0.05M tris-HCl, pH 8.5, 0.1mM EDTA). RNA was then re-purified along with 20 μg of mussel glycogen carrier by phenol:chloroform extraction and subsequent ethanol precipitation and resuspended in 7 μl water. The dephosphorylated RNA was then treated with tobacco acid pyrophosphatase (TAP), to remove the 5'-cap from the mRNAs, by incubation with 0.5 U of TAP in 10 μl TAP Buffer (0.05M sodium acetate, pH 6.0, 1 mM EDTA, 0.1% β-mercaptoethanol, 0.01% Triton X-100) at 37°C for 1 h. Afterwards the RNA was repurified and precipitated as described above. Next, the GeneRacer oligo-cap anchor RNA oligonucleotide (5'-CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA-3') was ligated to the 5'-end of the decapped mRNAs by adding the resuspended RNA (7 μl) to 250 ng of the RNA oligo as a dried residue in a microcentrifuge tube and incubating with 5 U of T4 RNA ligase in 10 μl T4 RNA Ligase Buffer (33 mM Tris-Acetate, pH 7.8, 66 mM potassium acetate, 10 mM magnesium acetate, 0.5 mM DTT) at 37°C for 1 h. The RNA was repurified and precipitated as described above and resuspended in 16 μl water.

Poly(A)+ RNA isolation

In order to remove non-mRNA species (for example, rRNA), poly(A)+ RNA molecules were isolated using Poly(A)Purist-MAG (Ambion) magnetic beads as directed. The oligo-capped RNA (in 16 µl water) was mixed with 16 µl 2x Binding Solution before adding 10 µg of washed Oligo(dT) beads and mixing by inversion. The beads and RNA were incubated for 5 min at 65 – 75°C and then for 45 min at room temperature with gentle agitation. The beads, now binding the poly(A)+ RNA were captured using a magnetic stand and washed twice with Wash Solution #1 and twice with Wash Solution #2. The poly(A)+ RNA was removed from the beads by adding two successive 100 µl aliquots of THE RNA Storage Solution. The poly(A)+ RNA, now in THE RNA Storage Solution was subject to an ethanol precipitation with 20 µl of 5 M ammonium acetate, 20 µg glycogen and 550 µl ethanol and incubation at – 20°C for at least 1 h. After centrifugation, the pellet was resuspended in 10 µl water. The poly(A)+ RNA isolated from this first round of poly(A)+ isolation was then subjected to a second, identical poly(A)+ isolation, including use of the glycogen carrier in the ethanol precipitation as this is lost in the poly(A)+ isolation steps, and final resuspension in 10 µl water.

First strand cDNA synthesis

The oligo-capped poly(A)+ RNA from the double isolation was then used for reverse transcription. The double-isolated poly(A)+ RNA (10 µl) was mixed with 1 µl reverse transcription primer 9200 (50 µM) and 1 µl dNTP mix (10 mM) and incubated at 65°C for 5 min, incubated on ice for 1 min and mixed with 4 µl of 5x First Strand Buffer (1x = 50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 0.01M DTT), 1 µl of 0.1 M DTT, 1 µl of RNase Out (40 U), and 1 µl (200 U). The assembled reaction was incubated

at room temperature for 5 min and then at 50°C for 1 h. After inactivating the enzyme at 70°C for 15 min, 2 U of RNase H was used to degrade the RNA. This provided us with PCR-ready single stranded cDNA. Clontech's Advantage 2 PCR Kit, a touchdown PCR method, was used to amplify the cDNA. For each 50 µl reaction, 1 µl of first strand cDNA, as described above, was mixed with 5 µl of 10x Advantage 2 PCR Buffer (400 mM Tricine-KOH, 150 mM KOAc, 35 mM Mg(OAc)2, 37.5 µg/ml BSA, 0.05% Tween 20, 0.05% Nonidet-P40), dNTPs (0.2 mM each), 0.1 ug of each primer and 1 µl 50x Polymerase mix (50% Glycerol, 15 mM Tris-Hcl, 75 mM KCl, 0.05 mM EDTA, 1.1 µg/µl TaqStart Antibody, TITANIUM TaqDNA Polymerase).

PCR amplification of cDNA

Initially, two separate PCR reactions were done: SL-PCR using an SL primer, or oligo-cap PCR using the GeneRacer 5' Primer as the forward primer, and both using the same reverse primer (oligo 9038, Appendix I). The thermocycler program was 94°C for 1 min, [94°C for 30 s, 57°C for 30 s, 68°C for 30 s] x25 cycles, 68°C for 2 min.

Because 25 cycle amplification of the initial cDNA product appeared to produce heteroduplexes involving primer-multimer 5'-RACE products we elected to initially amplify through a small number of PCR cycles of 94°C for 1 min, [94°C for 30 s, 57°C for 30 s, 68°C for 30 s] x5 cycles, 68°C for 2 min to produce a population of perfect duplex DNA for subsequent size-separation step by gel filtration column chromatography. A 30 µl aliquot of the 5 cycle double-stranded DNA was subsequently subjected to size-exclusion gel filtration chromatography on a ~1 ml Chroma Spin +TE-400 (ClonTech) column using gravity flow, and 18 one drop (~ 30 µl) fractions were collected. To detect the presence of DNA, 1 µl aliquots of each column fraction were

used as template DNA for PCR amplification under same conditions as used above but in final reaction volumes of 10 µl and assessed by electrophoresis on a 1% agarose gel. The first fraction(s) with detectable DNA, corresponding to the excluded volume of the column, were then utilized as the template for a new 50 µl 25 cycle PCR amplification as described above. The heterogeneous DNA products from this PCR reaction were then isolated following gel electrophoresis, cloned into pCR4-TOPO (Invitrogen) and plasmid DNAs from 8 randomly selected white clones were purified (see Standard plasmid cloning protocol). Insert sizes were verified by restriction digestion with EcoRI, which cleaves at two sites flanking the cloning site of pCR4-TOPO, and inserts were sequenced using the T3 sequencing primer.

Preparation of 5'-RACE libraries

Six 5'-RACE oligo-capped cDNA amplicon libraries representing four species were prepared for high-throughput sequencing. RNA from each species/stage was oligo-capped, reverse transcribed, subjected to a 5 cycle PCR amplification and subjected to gel filtration column chromatography to isolate the excluded volume fractions as described above. Then a 30 cycle 50 µl oligo-cap PCR on the best fraction(s) using the same parameters as outlined above was done using the 454 adapter primers which included a unique 10-base MID (Multiplex Identifier) barcoding sequence (Appendix I and Figure 10). The heterogeneous DNA products from each PCR amplification were then gel-isolated, cloned, and plasmid DNAs from 8 randomly selected white clones generated as described above (see Standard plasmid cloning protocol). Insert sizes were verified by EcoRI digestion and inserts were sequenced using the T3 sequencing primer.

The six amplified cDNA libraries: *Branchiostoma* adult (MID-2), *Branchiostoma* embryo (MID-3), *Saccoglossus* embryo (MID-7), *Strongylocentrotus* embryo (MID-10),

Petromyzon adult (MID-4) and *Petromyzon* embryo (MID-11), were provided to the McGill University and Genome Quebec Innovation Centre as samples of ~600 ng PCR products in 20 µl of 10 mM Tris-HCl, pH 8.5.

454 high-throughput sequencing

Before sequencing, each library was assessed by the McGill University and Genome Quebec Innovation Centre by Bioanalyzer gel electrophoresis (Appendix VI). An Ampure purification to remove small products was done on the MID-2, MID-3, MID-7 and MID-10 libraries. Libraries were pooled at equimolar proportions for each species, as follow: 25% MID-7, 25% MID-10, 20% MID-3 and 5% MID-2, 20% MID-11 and 5% MID-4. The pooled sample was then subjected to a second round of Ampure purification and 3 rounds of DNA-rich bead enrichment, which were required to generate sufficient DNA-rich beads, before running a half plate sequencing reaction on the Roche 454 GS-FLX Titanium instrument.

454 bioinformatics

In collaboration with Dr. Ken Dewar and Jessica Wasserscheid from the McGill University Genome Quebec Innovation Centre a bioinformatics analysis pipeline of the reads from each species was established.

The first step was to pool the amplicon and shotgun instrument interpretation pipeline reads for each library. To prepare a non-redundant set of pooled reads the shorter of the two read versions for any readname appearing twice, (i.e. originating from both pipelines) was discarded (usually that from the shotgun pipeline).

The second step was to trim the sequences of forward and reverse primer motifs. Many reads contained multiple partial copies of the forward primer. We also observed

that long amplicon pipeline reads invariably exhibited a double-base calling phenomenon starting from about 430 bases into the sequence. Thus in trimming the reads the reverse primer and its common variants (Table 1) were first removed. Then, long sequences were truncated to 400 nucleotides in order to eliminate the region of double-base calling. Following this, the forward primer and its common variants (Table 1) were trimmed; including those found within reads, such as would be the case for primer-multimers. In addition, in order to facilitate subsequent alignments, any A residues present at the junction between the oligo-cap primer and the mRNA 5'-sequence were trimmed (Figure 3). The oligo-cap ribonucleotide ends in the motif GAAA before the 5'-end of the mRNAs begin. Although many reads had three A residues, a significant proportion of reads had from 1-9, and, exceptionally, more than 9 A residues at this site. It is possible that much or all of this variability is due to the 454 system's difficulty in reading homopolymers. Alternatively, it is also possible that the 5'-end of these mRNAs is one or more A residues.

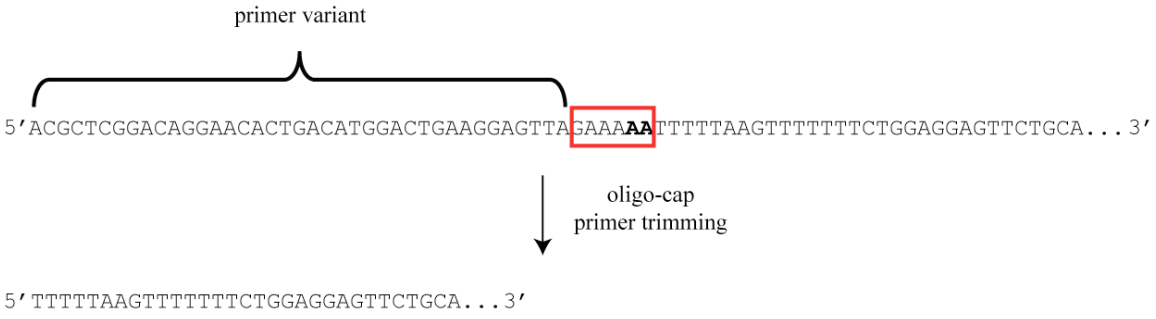


Figure 3: Example of forward primer trimming. The top line shows the 5'-sequence of a 454 read as reported by the Roche 454 GS-FLX Titanium instrument. The sequence begins with a variant of the forward primer, recognized by manual inspection, which is trimmed. In addition, GAAA, originating from the oligo-cap RNA oligonucleotide is trimmed. Finally any additional A residues following the GAAA motif (bold) are also trimmed. The bottom line shows the sequence of the trimmed 5'-end of mRNA ready to be examined for the possible presence of a SL.

Finally, any trimmed read that was less than 50 nucleotides in length was discarded.

Table 1: Primer motifs and common variants trimmed in bioinformatics pipeline

Forward primer*: Common Forward primer variants*:	5' GGACTGAAGGAGTAG 3' 5' GGACTGAAGGAGTCG 3' 5' GGACTGAAAGGAGTAG 3'
Reverse primer: Common Reverse primer variants:	5' ATATGACTCCTCTGAGGTAGAATT 3' 5'ATATGACTCTCTGAGGTAGAATT 3' 5' ATATGACTCCTCTGAGTAGAATT 3'

*These forward primer motifs are 13 nucleotides (+/- 1 nt) from the 5'-end of the Roche 454 GS-FLX Titanium instrument-reported reads. Forward primer trimming was from the first base of the reported sequence up to the last G in the primer motif.

The third step in this bioinformatics analysis of each MID library was a preliminary genome alignment using BLAT in order to identify, for subsequent removal, less common primer sequence variants that had escaped the above primer trimming operations. The genomes used in this and all subsequent bioinformatics analyses are listed in Table 2. The standalone BLAT version 33 (found at http://users.soe.ucsc.edu/~kent/src/) was used with parameters set to reproduce the web-based UCSC BLAT results: tileSize=11, stepSize=5, minIdentity=0, minScore=0, with all other parameters kept at default. In many cases, variant primer motifs that had escaped the bioinformatics trimming step resulted in read alignments starting from internal positions. All reads aligning from internal positions were visually inspected and recognizable variant primer motifs were trimmed (see Figure 3). Any reads less than 50 nucleotides after trimming were discarded.

Table 2: Genome assemblies used for bioinformatic analysis

Species	**Assembly Version**	**URL**
Branchiostoma	braFlo1 (Mar. 2006)	http://hgdownload.soe.ucsc.edu/downloads.html#lancelet
Saccoglossus	Skow_1.1 (01-Jul-2009) assembly (scaffolds fasta file)	http://www.hgsc.bcm.tmc.edu/ftp-archive/Skowalevskii/fasta/Skow_1.1/
Strongylocentrotus	strPur2 (Sep. 2006)	http://hgdownload.soe.ucsc.edu/downloads.html#urchin
Petromyzon	petMar1 (Mar. 2007)	http://hgdownload.soe.ucsc.edu/downloads.html#lamprey

Once processed to eliminate primer-dimers, primer-multimers and short reads the number of reads in all libraries were significantly reduced. The next step of the bioinformatic analysis of the MID libraries was a definitive genomic alignment using BLAT to generate an output table for each read set with the following information: 1) read name, 2) sequence, 3) length of sequence, 4) top score, 5) alignment query start (base at which read starts aligning to the genomic sequence), 6) alignment query end (last base that aligns to the genomic sequence), 7) % identity (# identical nucleotides/length of query read), 8) alignment polarity (+ or – strand), 9) target (chromosome or scaffold on which the sequence is found), 10) alignment genome start position, 11) alignment genome end position, 12) genome span and 13) sequence of the 50 bases directly upstream of the genome start position. Columns 4) through 13) were repeated for the 2nd, 3rd, 4th and 5th BLAT hits and were included on the table.

In manually analyzing these reads and their top hits to the genome, it became apparent that a significant minority had very low % identity scores (# identical nucleotides/length of query read) with only ~ 20 bases of their sequence matching to the

genome. These were likely random matches to the genome and not representing true genes. By systematically using sequences from the different libraries as BLAST queries against their respective EST databases we found that reads with % identity scores of ≥ 55% appeared to align at the 5'-ends of real ESTs. In order to avoid accidentally discarding a read that had a score less than 55% because it had a long spliced leader sequence at its 5'-end, the reads that failed the 55% cutoff were subjected to a second round of quality control. If a read had 30 bases or more aligning to the genome this was considered too long to occur by chance, and thus the read was not discarded, even if the % identity score was < 55%. There were only a few (2-5) examples of such reads from each species and most often they appeared to represent 5' or 3' incomplete genes, possibly due to incomplete genome assembly. Thus all reads whose top hit % identity score was lower than 55% and whose alignment length was less than 30 nt were discarded.

The last column to be made for these tables was a gene counter. For identification of a common 5'-end on different mRNAs it was imperative to confirm the reads were from different genes and not simply the same mRNA represented multiple times. The gene counter algorithm was based on genomic location. Basically if two reads had overlapping positions in the genome they were considered to derive from the same gene. To generate this table required first that the reads be separated into those from the + strand and those from the – strand, and then for each table sorted by chromosome/scaffold. The + strand reads were then sorted in ascending order by their genome start position while the – strand reads were sorted by their genome end position in descending order, thus both the + and – strand tables would have the most 5' located read in the top row. This top row read is, in each table, assigned to Gene 1 and then the second row read is considered as to whether it too derives from Gene 1, or whether it

defines a new gene, Gene 2. If it maps to a different chromosome/scaffold, or if the 5'-end of read 2 maps downstream of the 3' end of the read 1, then read 2 is assigned to Gene 2. The total number of different mRNA species in each read set could then be determined by adding the gene count from the + and – strands together.

Statistical analysis

The probability that SL *trans*-splicing was occurring, but was missed because of sampling was calculated by addition of the probability of not observing any SL *trans*-spliced mRNA species and the probability of observing only one SL *trans*-spliced mRNA species (since without the 5'-end being common to another mRNA it could not be recognized as a candidate SL). Thus, the probability (p) of not finding SL *trans*-splicing given a gene/mRNA species sample size of (n) and a fraction (F) of genes that undergo *trans*-splicing is, $p = (1 - F)^n + nF(1 - F)^{n-1}$.

CHAPTER 3 – ANALYSIS OF THE DISTRIBUTION OF SL *TRANS*-SPLICING IN THE DEUTEROSTOMES

INTRODUCTION

In order to explore the evolutionary history of SL *trans*-splicing we decided to examine mRNA 5'-ends of four deuterostome species. These four species, *Strongylocentrotus*, *Saccoglossus*, *Branchiostoma* and *Petromyzon* were ideal candidates to examine for SL *trans*-splicing as they are key deuterostome groups whose *trans*-splicing status is unclear and for which there is good genome sequence information. Although Douris et al. (2010) categorize all four species as not utilizing SL *trans*-splicing, their analysis, based on conventional EST data, would have missed any short SLs such as the 16 nt SL of *Ciona*. The only way of definitively determining the SL *trans*-splicing status of a species is to examine 5'-RACE products looking for a common 5'-end on a variety of different mRNA species.

5'-RACE methods

There are several different commonly used 5'-RACE methods, including dC tailing, template-switching (SMART) and oligo-capping. The dC tailing 5'-RACE method, available commercially through Invitrogen, is based on reverse transcription of the RNA followed by the addition of a homopolymer C tail on the 3'-end of the first strand cDNA using terminal deoxynucleotidyl transferase (TdT) and dCTP (Frohman, Dush, & Martin, 1988; Loh et al., 1989). The resulting C-tailed cDNA can be amplified by using an anchor primer recognizing the C tail and a reverse primer. The SMART 5'-RACE system (Clontech), similarly adds bases at the 3'-end of first strand cDNA, but does so through terminal transferase activity of the reverse transcription enzyme (Matz et al., 1999; Zhu et al., 2001). An anchored oligonucleotide including a 3'-tail

complementary to the additional bases added to the first strand cDNA is also present in the reaction and the reverse transcriptase enzyme is then able to switch templates and extend the second strand (Matz et al., 1999; Zhu et al., 2001). Oligo-capping (Invitrogen), is based on RNA ligase-mediated ligation of an arbitrary RNA oligonucleotide at the 5'-end of full-length capped mRNA molecules (Maruyama & Sugano, 1994). Reverse transcription generates the first strand cDNA with sequences complementary to the oligo-cap primer at the 3' end, which is recognized and utilized by an anchored primer in the PCR amplification (Maruyama & Sugano, 1994).

As mentioned above, we wish to examine the 5'-ends of the mRNA population of different species. As oligo-capping is the only method that will selectively amplify 5'-complete molecules due to the required participation of the 5'-cap structure, we chose this 5'-RACE method for our analysis.

To generate an oligo-capped cDNA library representing many different genes we could either use oligo(dT) priming to obtain full-length molecules or random priming to obtain molecules with variable 3'-ends (Suzuki et al., 1997; Suzuki et al., 2002). Since our interest is exclusively at the 5'-ends there is no inherent need for the 3'-end of the mRNA molecules. Moreover, oligo(dT) primming would restrict our libraries to only shorter transcripts (i.e. those short enough that the enzyme is able to reverse transcribe from the poly(A) tail to the 5'-end without falling off). Therefore, we chose to use random priming (Suzuki et al., 2002). Hence our cDNA libraries would include the 5'-ends of all lengths of transcripts, even very long ones.

High-throughput sequencing

The plan was to sequence these 5'-RACE products by a "next-generation" high throughput sequencing approach. The two options to be considered for high throughput

sequencing were Illumina or Roche 454 systems. Because of the comparatively short read length of Illumina, 100 nucleotides, vs. 450 or more nucleotides for 454 (Mardis, 2008), and the requirement that we sequence entirely through the whole length of any 5' SL sequence and into the unique-sequence segment of each mRNA, the 454 method seemed better suited. The average sequence length is 450 nucleotides, allowing the use of BLAST to confirm the identity of the originating species. As a second way to identify the originating species of each read, the PCR primers used on the oligo-capped first strand cDNA were engineered to include several nucleotide changes specific to each of the four species thus creating a barcode (Parameswaran et al., 2007).

RESULTS

Elimination of artifactual primer-multimer 5'-RACE products

In order to validate the oligo-capping method as suitable for detecting short SLs, a preliminary 5'-RACE experiment was completed using total RNA from adult *Ciona intestinalis* body wall muscle previously isolated in the lab by Trizol extraction. The RNA was processed following the GeneRacer RLM-RACE (Invitrogen) protocol (see Chapter 2). Briefly, the RNA was first treated with bacterial alkaline phosphatase (BAP), which removes the exposed 5' phosphates from any non-capped RNA, including any mRNA 3'-fragment possibly generated by mRNA breakage before or during extraction. The RNA was then treated with tobacco acid pyrophosphatase (TAP), removing the cap structure from mRNA and leaving a 5' monophosphate only on RNA molecules deriving from the 5' terminus of capped RNAs (e.g. mRNA). T4 RNA ligase was then used to ligate an arbitrary-sequence anchor RNA oligonucleotide to these mRNA 5'-monophosphate full-length molecules (or 5'-segments). The poly(A)+ RNA is then isolated using Poly(A)Purist Magnetic beads (Ambion) to remove any contaminating

rRNA. The poly(A)+ RNA was then reverse transcribed using a reverse transcriptase and random primers (random hexamers linked 3' to an arbitrary anchor sequence) to obtain first-strand cDNA products representative of all RNA molecules present. Subsequently the RNA strands were digested with RNase H, and an oligo-cap anchor-specific forward primer and a 3'-anchor-based reverse primer used to amplify the cDNAs through PCR.

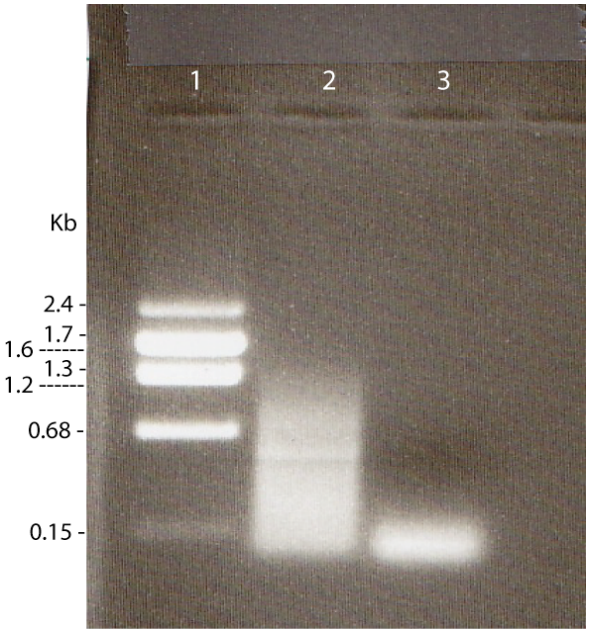


Figure 4: SL PCR and oligo-cap PCR products from *Ciona* body wall muscle RNA. Poly(A)+ RNA isolated following ligation of the oligo-cap anchor RNA oligonucleotide was reverse transcribed to give first strand cDNA. This cDNA was then amplified by two methods of PCR: SL-PCR using an SL primer, or oligo-cap PCR using the GeneRacer 5' Primer as the forward primer. Both PCR reactions used the same reverse primer. Products were analyzed by electrophoresis on a 1% agarose gel. Lane 1: DNA size ladder, Lane 2: SL PCR products, Lane 3: oligo-cap PCR products.

Initially two PCR amplifications of the *Ciona* oligo-cap cDNA were done, one with an oligo-cap primer, termed GeneRacer 5' Primer (provided by Invitrogen) and the other with a primer, termed SL primer (oligo 9015, 23 nt, see Appendix I), containing at it's 3'-end the SL sequence as a positive control to verify that we did indeed have 5' complete cDNAs, including the SL.

The size distribution of the products from the two PCR amplifications was assessed by gel electrophoresis. From a previous high-throughput analysis of *Ciona* SL *trans*-spliced mRNAs using the same primers it was expected that the SL PCR products would have a size distribution of 100 – 850 bp, with a mean of 403 bp (Matsumoto et al., 2010). Indeed, in this experiment, the SL PCR products had an

appropriate size distribution from 100 – 1000 base pairs, with the majority being around 500 bp. However, the oligo-capped PCR products from amplification with the GeneRacer 5' Primer all appeared to be around 150 base pairs (Figure 4).

In order to validate that these ~150 bp PCR products were indeed derived from the cDNA and were not primer-dimer products, a second PCR experiment was done. Included in this set of PCR amplifications was; 1) a repeat of the initial oligo-capped primer PCR amplification with the GeneRacer 5' primer and 2) a negative control, including everything except the cDNA in the reaction. The results confirmed that the occurrence of the ~150 bp products seen from GeneRacer 5' primer PCR was dependent on the presence of cDNA (Figure 5).

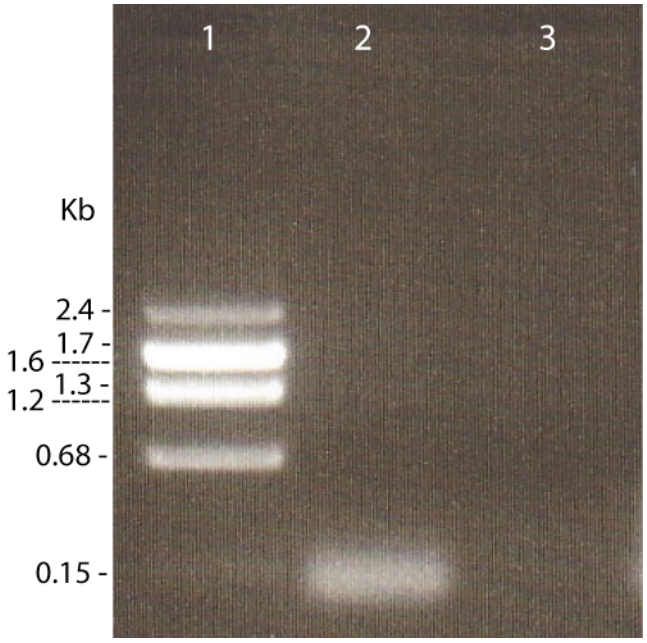


Figure 5: Amplification of the ~150 bp oligo-cap PCR products are dependent on the presence of template cDNA. Oligo-cap PCR using the GeneRacer 5' Primer was done with (Lane 2) or without (Lane 3) cDNA template. Products were analyzed by electrophoresis on a 1% agarose gel.

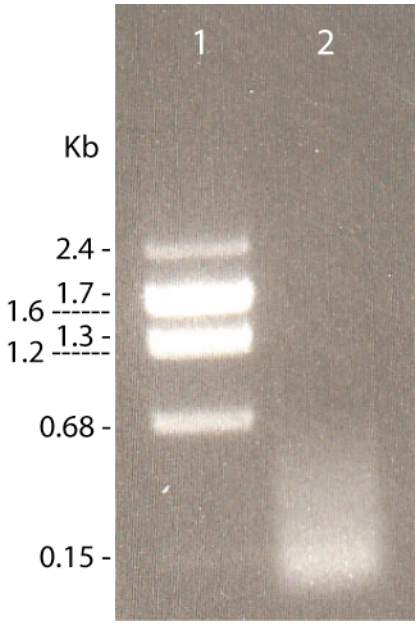


Figure 6: Analysis of *Ciona* body wall muscle RNA oligo-cap PCR with the GeneRacer 5' Nested Primer. Products were analyzed by electrophoresis on a 1% agarose gel. Lane 1: DNA size ladder, Lane 2: Oligo-cap PCR with GeneRacer 5' Nested Primer.

In an attempt to obtain a better size distribution of oligo-capped PCR products, including larger molecules, a nested oligo-capped PCR primer (provided by Invitrogen) was utilized. This alternative oligo-cap primer corresponds to a sequence near the 3'-end of the oligo-cap RNA oligonucleotide. It is termed the GeneRacer 5' Nested Primer and is generally used in a second amplification after a first round of amplification with the GeneRacer 5' Primer. However, the GeneRacer 5' Nested Primer was tried directly on the unamplified cDNA in case the initial GeneRacer 5' Primer did not give good products due to unfavorable interactions with the reverse primer. It appeared that using the GeneRacer 5' Nested Primer gave a better size-distribution of products; ~150 bp products remained the majority, however, now the visible products ranged up to 700 bp (Figure 6). To elucidate the character of these PCR products they were investigated by cloning and sequencing. DNA products from the nested PCR amplification in the size range of 300 – 700 bp was cut out of the gel and extracted. The DNA was then used for TA cloning. The cloning reaction was then transformed into high competency DH5α cells and plated on LB agar with ampicillin. The same protocol was followed for SL PCR products. Plates were provided to the McGill University and Genome Quebec Innovation Centre for colony picking and culture growth from which the plasmid DNAs were isolated and sequenced using the T3 sequencing primer. Three

separate plates were analyzed: One contained colonies from the GeneRacer 5' Nested Primer PCR amplification, from which seventy-four were picked and sequenced; one contained colonies from the SL PCR, from which twenty were picked; and one contained colonies carrying pBluescript SK II + vector which, like the pCR4-TOPO vector includes the T3 promoter sequence, from which two clones were picked.

The sequencing results from the pBluescript SK II + and the SL PCR colonies were exactly as predicted. Both pBluescript SK II + sequences did indeed align perfectly to the known pBluescript SK II + sequence. All 18 sequences from the SL PCR clones (originally 20 clones, but 2 failed internal quality control filters according to the Genome Centre) except one, with a short 65 bp insert, contained the SL sequence and gave high quality alignments to the *Ciona* genome and/or ESTs (Appendix III). The GeneRacer 5' Nested Primer PCR sequences were also expected to have aligned to the *Ciona* genome and/or EST data, however this was often not the case. Of the 74 sequences obtained, only 14, representing 11 different mRNA species appeared to be authentic cDNAs based on high quality alignments with the *Ciona* NCBI EST database (See Appendix III). Two additional sequences appeared to be cDNAs but did give any high quality alignment (data not shown). The other 58 consisted only of primer-related sequences. Of the sequences that contained only primers, the GeneRacer 5' Nested Primer was observed in multiple tandem copies (from 2-11 times). Often, unexpectedly, these repeats contained additional sequence belonging to the 5'-end of the oligo-cap anchor RNA oligonucleotide sequence (i.e. upstream of where the PCR primer hybridized) (See Figure 7)

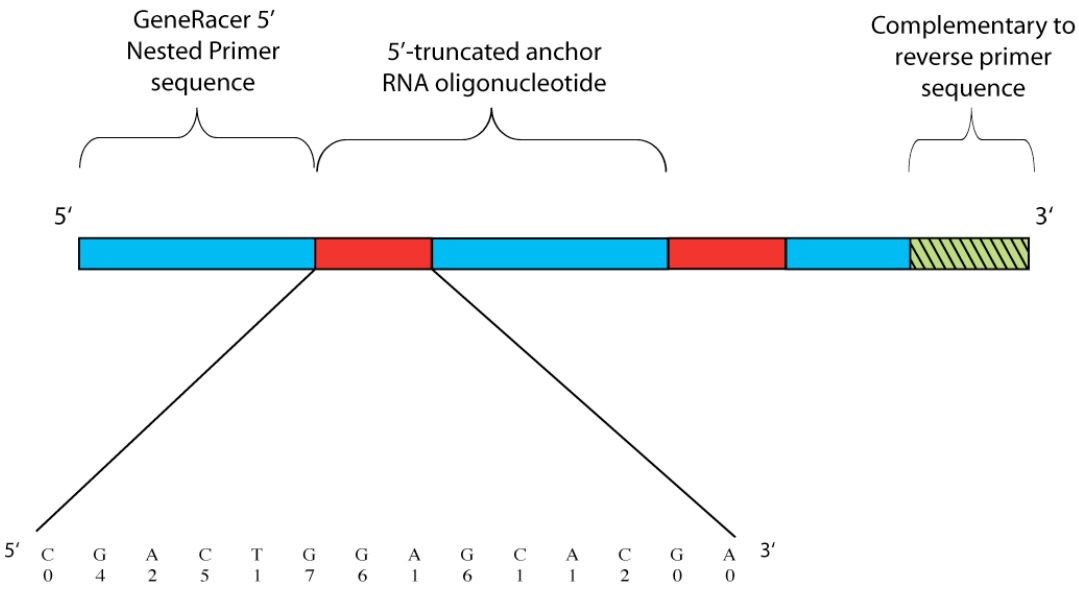
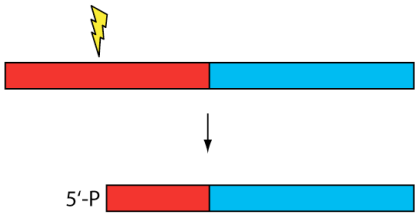


Figure 7: General configuration of oligo-cap 5'-RACE primer-multimer PCR products and length of additional sequence belonging to the 5'-end of the anchor RNA oligonucleotide sequence. General primer-multimer product contains at its 5'-end the full GeneRacer 5' Nested Primer sequence, then a 5'-truncated anchor RNA oligonucleotide sequence, (repeated X times, only shown once in figure), including 3-13 nucleotides of the 14 nt RNA oligonucleotide sequence upstream of the motif represented by the GeneRacer 5' Nested Primer. This upstream sequence is shown below the schematic, with numbers showing the observed frequencies of occurrences of 5'-truncation sites in 36 randomly chosen cases. The most downstream copy of the RNA oligonucleotide sequence is also truncated at the 3'-end where it joins to the reverse primer, presumably reflection random priming of reverse transcription within a copy of the RNA oligonucleotide.

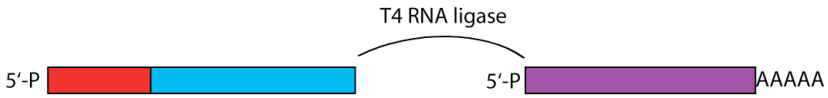
One possible interpretation is that such molecules could derive from multimerized RNA ligase products in which a full-length anchor RNA oligonucleotide has been ligated to an anchor RNA oligonucleotide 3'-fragment carrying a 5'-phosphate (termed a 5'-truncated RNA oligonucleotide). RNA oligonucleotide breakage/cleavage could occur before or during the RNA ligase reaction. (Figure 8).

Figure 8: Proposed mechanism for generation of primer-multimer products.

(1) Unexpected cleavage of minority of full-length anchor RNA oligonucleotides generating 5′-truncated 5′-phosphorylated RNA oligonucleotides

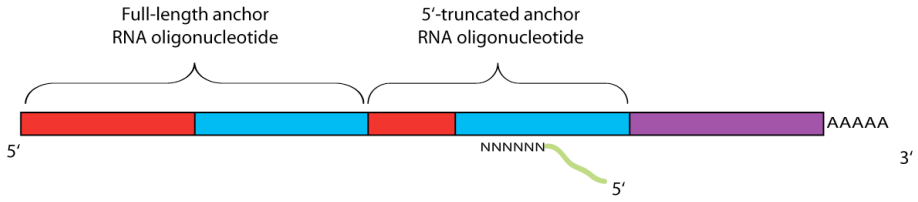


(2) 5′-truncated anchor RNA oligonucleotides may be ligated to decapped mRNAs

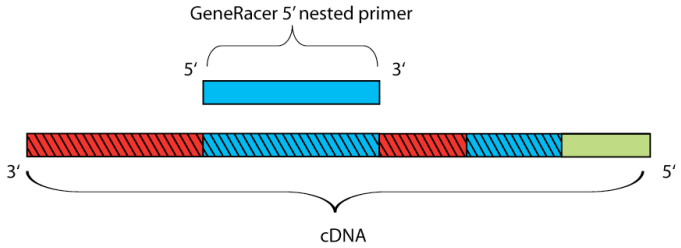


(3) 5′-truncated anchor RNA oligonucleotide-mRNA molecules are further ligated to a full-length anchor RNA oligonucleotide

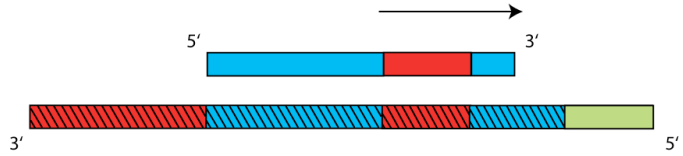
(4) cDNA synthesis: reverse-transcription primer random hexamer sequence anneals to the anchor RNA oligonucleotide and primes the cDNA



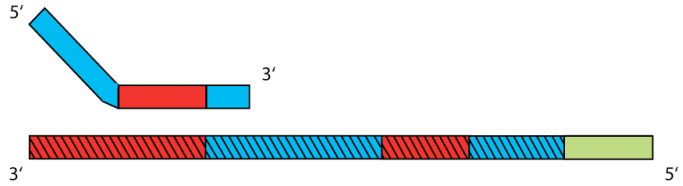
(5) GeneRacer 5′ nested primer anneals to the upstream (full-length) copy of the complementary sequence



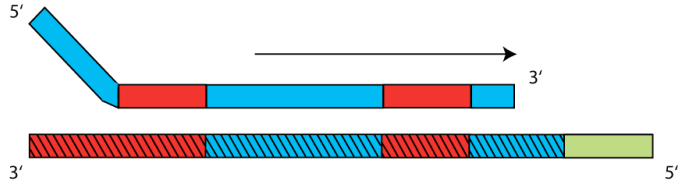
(6) Initial extension of the primer by Taq DNA polymerase



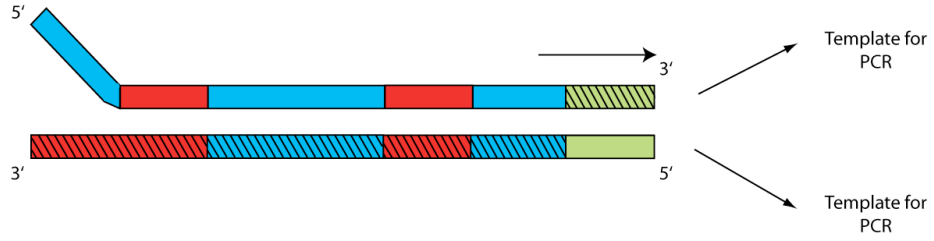
(7) Slippage of the nascent plus strand product



(8) Extension of the slipped nascent strand by Taq DNA polymerase



(9) Further extension of the slipped nascent strand by Taq DNA polymerase creating a new cDNA template for PCR



In these 58 sequences there did not appear to be any pattern in how much of the 5'-end of the oligo-cap RNA oligonucleotide sequence was included, meaning there appeared to be no preferred cleavage site. However, generally, if there were multiple copies of the primer in a sequence, the multiple copies all contained precisely the same length of upstream 5'-end of the RNA oligonucleotide. The homogeneity of the length of the "extra" RNA oligonucleotide upstream sequence segment among the repeat units within a single PCR product could result from additional expansion of the repeat copy number during the PCR amplification by a slippage mechanism (see Figure 8). An initial extension of the PCR nested primer by Taq DNA polymerase could occur from the full-length 5'-terminal RNA oligonucleotide sequence into the downstream 5'-truncated RNA oligonucleotide sequence. Then, slippage of this nascent plus strand product could occur, and with further extension of the slipped nascent plus strand by Taq DNA polymerase these repeated products could be made. The average size of these primer-multimer products was 162 bp (range 45 bp to 405 bp), corresponding to the ~150 bp size of the major product seen in the gel analysis of the PCR products. Note that prior to cloning DNA products of apparent size 300 – 700 bp were recovered following gel electrophoresis. This discrepancy between electropohretic apparent size and ultimate sequence length becomes an important theme. In all 74 of the sequences, the reverse primer was always present once per sequence and sometimes began at various sites within the most downstream forward RNA oligonucleotide motif (Figure 9). This could be explained by random-priming of the reverse transcription primer, within RNA oligonucleotide multimers, as in Figure 8, step 4.

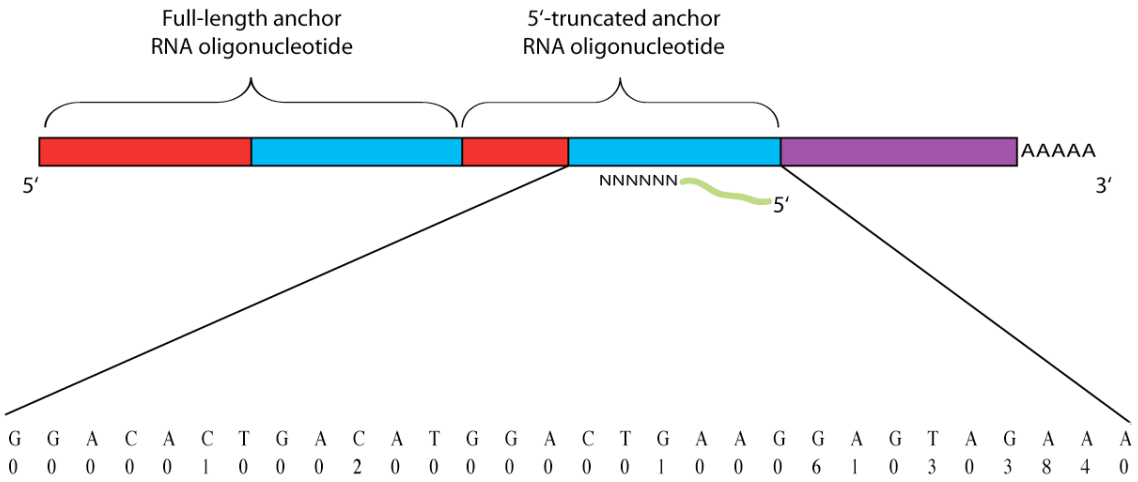


Figure 9: Distribution of reverse transcription random-priming sites within the RNA oligonucleotide. The sequence of the 30 nucleotides at the 3'-end of the RNA oligonucleotide is shown below the schematic. The associated numbers show the distribution among 29 randomly chosen cases, of the junction between the RNA oligonucleotide sequence and the sequence complementary to the reverse primer. Because the reverse transcription primer contained a random hexanucleotide sequence 3' of the anchor sequence, the actual site of cDNA priming are 6 nt to the left of the indicated nucleotides.

These findings in our preliminary experiments with the "control" *Ciona* body wall muscle RNA sample informed our oligo-cap 5'-RACE analysis of "test" species, which we began with the *Petromyzon* adult muscle RNA sample. In an attempt to remove any primer-multimer products before cloning and sequencing we decided during our initial experiments with adult *Petromyzon* muscle RNA to include a double poly(A)+ RNA isolation step in case the primer-multimer RNA templates could be ligated RNA oligonucleotide multimers unconnected to poly(A)+ mRNA molecules that were not entirely removed by a single poly(A)+ RNA isolation step. Incorporating the two poly(A)+ RNA isolations did appear to improve the distribution of the PCR products as the range of their sizes was larger, 100 – 2000 bp, than was observed with a single poly(A)+ RNA isolation however the ~150 bp band was still clearly visible (Figure 10).

As an additional measure to eliminate short primer-multiple products, a higher size cutoff was used; products of >500 bp apparent length were gel-recovered prior to cloning. The PCR products were TA cloned and plasmid DNA from 8 randomly picked clones was isolated and digested with EcoRI to release the insert. Surprisingly, agarose gel electrophoresis showed 7 of the 8 clones had the canonical ~150 bp insert, suggesting that these were primer-multimer products, similar to the average size of the multiple primer products seen previously in the *Ciona* oligo-capping experiment.

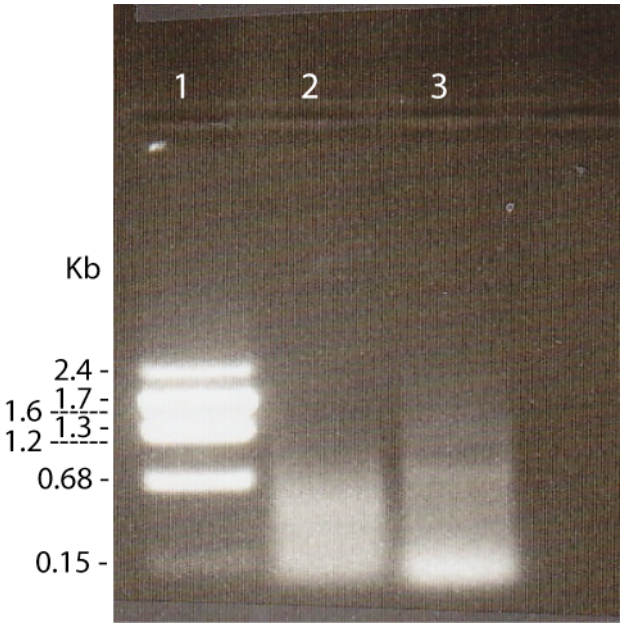


Figure 10: Two serial poly(A)+ RNA isolations improve the size distribution of Petromyzon muscle RNA oligo-cap 5'-RACE PCR products. After ligating the oligo-cap RNA oligonucleotide to the RNA, poly(A)+ RNA was isolated using Poly(A)Purist magnetic beads prior to reverse transcription and oligo-cap PCR with the GeneRacer 5' Nested Primer. Products were analyzed by electrophoresis on a 1% agarose gel. Lane 1: DNA size ladder, Lane 2: One poly(A)+ isolation, Lane 3: Two sequential poly(A)+ isolations.

Our hypothesis to account for the surprising cloning of short PCR product inserts from > 500 bp products recovered following agarose gel electrophoresis was that in the PCR a heteroduplex was being formed. When amplifying to high DNA concentrations, the short primer-multimer products could compete with bona fide primers during the annealing step, forming a heteroduplex with both ends completely double stranded and a large single stranded internal loop (Figure 11). During agarose gel electrophoresis such a heteroduplex molecule could appear as much larger than would ~150 bp perfect duplex

products. Then, in the cloning process the ~150 bp strand of the heteroduplex could perhaps be preferentially incorporated into the vector.

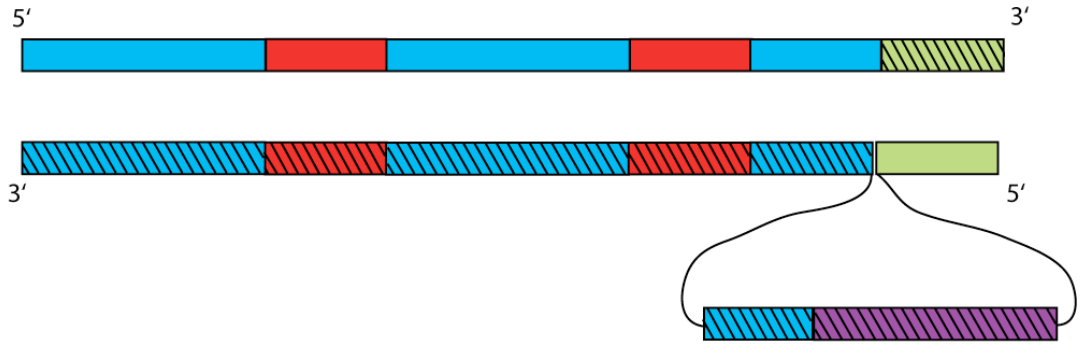


Figure 11: Heteroduplex molecule. The primer motif colour scheme is the same as in Figures 7-9. The two schematic molecules shown are the two strands of a heteroduplex molecule and are base-paired throughout their complementary sequences. In the case shown, where one strand consists of primer-multimers and the other strand includes a cDNA "insert", the latter will not be base-paired but will form a large single-stranded loop (not shown to scale). When amplifying to high DNA concentrations, the short multiple primer sequence (top) can compete with bona fide primers during the annealing step, forming a heteroduplex.

In order to try to prevent primer-multimer products from becoming the dominant PCR products we decided to subject early-cycle PCR products, when DNA concentrations would be too low to permit heteroduplexes forming during the annealing step, to gel filtration column chromatography (Chroma Spin 400) to remove any small (< 600 bp) products and then do a full PCR with the recovered > 600 bp DNA as the template. This new PCR strategy was tested with the adult muscle *Petromyzon* cDNA. After 5 cycles of amplification of the reverse transcription products with the GeneRacer 5' Nested forward primer and the 3'-anchor-based reverse (9038) primer, the PCR products were passed through the column to separate larger products (legitimate 5'-RACE cDNA products) from smaller products (primer-multimer molecules). The DNA content of column fractions, too low for direct detection, was assessed by 30 cycle amplification of an aliquot of each fraction with the same primers. This revealed the

cDNA elution profile shown in Figure 12. The appropriate fractions containing larger products (Figure 12, lanes 11 and 12 corresponding to column fractions 9 and 10) were subsequently amplified for 30 cycles using the GeneRacer 5' Nested forward primer and the 3'-anchor-based reverse primer.

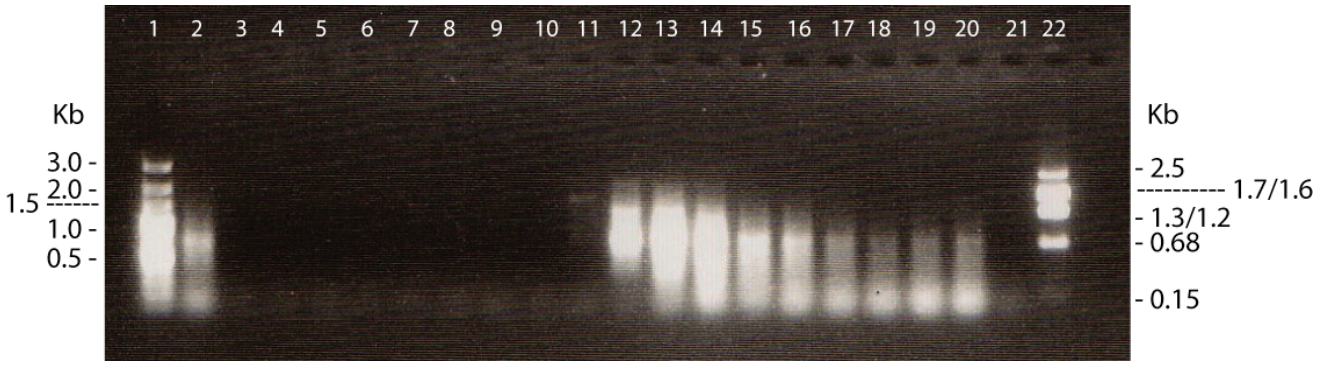


Figure 12: Gel filtration of early cycle 5'-RACE PCR products to separate larger cDNA molecules from smaller primer-multimer products. After 5 cycle PCR of *Petromyzon* oligo-capped cDNA with the GeneRacer 5' Nested Primer and 3'-anchor-based reverse primer, products were subjected to a gel filtration column. Chromatography column fractions were used as template DNA for 30 cycle PCR amplification with the same primers to reveal the DNA population present in each fraction. Lanes 1 and 22: DNA size ladders, Lane 2: Positive control PCR, PCR of non-chromatographed 5 cycle products, Lanes 3 – 20: 30 cycle PCR using different fractions, loaded in same order as eluted from the column, Lane 21: Negative control PCR (minus template DNA).

PCR products in the 500 bp – 1500 bp size range were gel extracted and analyzed by TA cloning and sequencing. All 8 clones tested had large inserts (with only one being < 450 bp) and sequencing confirmed none consisted of primer-multimers, with 6 aligning to *Petromyzon* ESTs (see Appendix IV, Section A, Rows 1-8). This validated the use of column chromatography of low cycle-number amplification products to remove the small primer-multimers.

This gel filtration step was included as a routine step in the oligo-capping 5'-RACE analysis of *Ciona*, embryo *Petromyzon*, embryo and adult *Branchiostoma*, embryo

Saccoglossus and embryo *Strongylocentrotus* samples. In each case, assessment by TA cloning and sequencing revealed inserts of at least 500 bp, many of which aligned with ESTs, exactly as the case was for adult *Petromyzon* above (see Appendix IV, Section A, Rows 9-48).

Introducing 454 high-throughput sequencing amplification and sequencing primer sequences into the 5'-RACE amplicon libraries

The next step was to do 454 high throughput sequencing of these 5'-RACE PCR libraries. Briefly, 454 sequencing involves annealing of individual single stranded DNA molecules to complementary sequence present on 28 μM beads (Margulies et al., 2005). These beads are then captured in droplets of PCR reaction mixture in an oil emulsion and PCR amplifies the DNA within each droplet creating on the order of 10 million bead-coupled copies of the DNA sequence (Margulies et al., 2005). The emulsion is then broken and DNA denatured to single-stranded DNA (ssDNA). The beads now carrying ssDNA are deposited into picolitre reaction wells and the DNA sequenced by introducing one dNTP and DNA polymerase every cycle with the 3'-end of the bead-coupled PCR primer sequence being extended until a different base is needed (Margulies et al., 2005). The amount of pyrophosphate released by nucleotide polymerization is assessed by sulfurylase-based light-generating assay (Ronaghi, Uhlen, & Nyren, 1998). The light generated in each picolitre reaction well is captured by a fiber optic camera system and the series of light flashes are interpreted in terms of DNA sequence (Ronaghi, et al., 1998; Margulies et al., 2005).

In order to provide 454 sequencing ready oligo-capped cDNA libraries to the McGill University Genome Quebec Innovation Centre it was necessary to introduce the forward and reverse 454 amplification primer sequences. I designed a set of 66-base

forward primers composed of the 454 amplification/sequencing primer elements, followed by the 4 base key used to calibrate light emission in terms of single A, G, C or T nucleotides, then a 10-base MID (Multiplex Identifier) barcoding sequence (which allows for samples from different libraries to be pooled together) and at the 3'-most end of the primer a template-specific sequence. In the case of our libraries, the template-specific sequence was the GeneRacer 5' Nested Primer sequence to allow for incorporation of these 454 adapter primers by PCR. Therefore the seven forward 454 adapter primers were identical, except for the 10-base MID barcode sequence (Figure 13). The 55 base reverse primer did not require a barcoding sequence and was identical for all libraries. It included the 454 emulsion PCR reverse primer sequence and a template-specific sequence, namely the 3'-anchor-based reverse primer sequence (originally part of the 9038 reverse transcription primer).

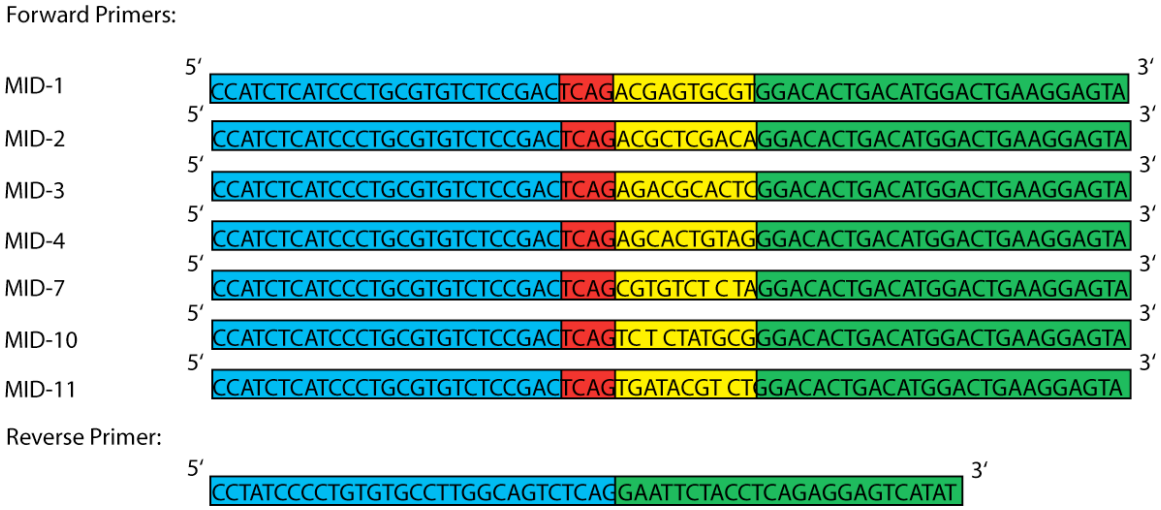


Figure 13: Basic structure of 454 amplification primers. The seven different forward primers used are shown. The reverse primer was identical for all libraries. Blue: 454 GS-FLX Titanium sequence (Primer A for Forward Primers, Primer B for Reverse Primer) sequence, Red: 4-base key, Yellow: Unique MID tag, Green: template specific GeneRacer 5' Nested Primer sequence (for Forward Primers), 3'-anchor-based reverse primer sequence (for Reverse Primer).

Initially the *Ciona* library, our control species known to utilize SL *trans*-splicing, was used to assess for any problems in the addition of the very long (55-66 bases) 454 adapter primers by PCR.

Using the 500 – 1500 bp adult *Ciona* cDNA library, that had been verified by TA cloning, as the template DNA, 454 adapter primers (MID-1 and the reverse primer) were added to the 5'- and 3'-ends of the *Ciona* cDNA population by an additional number of PCR cycles (see Figure 16). Wary of having too many plateau-phase PCR cycles during which heteroduplexes would form, 10 vs. 20 vs. 30 cycle PCR was assessed by gel electrophoresis to determine the minimum number of cycles required to obtain a cDNA population of at least 600 ng. This preliminary PCR indicated that visible products appeared between 10 and 20 cycles. To establish more precisely the minimum number of PCR cycles needed for the 454 adapter primer addition, 15 vs. 20 cycles was compared (Figure 14). The products were readily detectable after 15 cycles, but most interestingly, even though they were amplified from a gel-isolated template population of > 500 bp products these 454 adapter primer products appeared considerably smaller; ranging from lower than 400 bp down to ~ 100 bp. Moreover, there appeared to be an increase in the size of the PCR products between 15 and 20 cycles, and no increase in DNA amount (rather, an apparent decrease, perhaps reflecting the formation of single-stranded heteroduplex loops which would bind less ethidium bromide), suggesting again that heteroduplexes were forming sometime between cycle 15 and cycle 20.

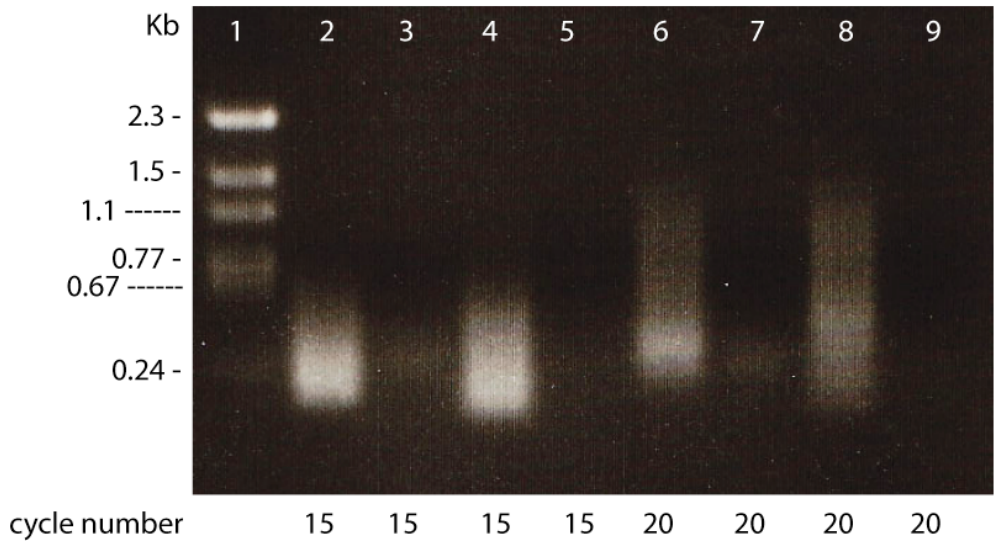


Figure 14: 15 cycle vs. 20 cycle oligo-cap PCR using original and 454 adapter primers on the 500 – 1500 bp adult *Ciona* 5'-RACE amplicon library. The increased apparent size of products after 20 cycles, suggests heteroduplex formation. Lane 1: DNA size ladder, Lane 2: 15 cycle PCR with GeneRacer 5' Nested Primer and reverse primer 9038, Lane 3: no-DNA control, Lane 4: 15 cycle PCR with 454 adapter primers, Lane 5: no-DNA control, Lane 6: 20 cycle PCR with original primers, Lane 7: no-DNA control, Lane 8: 20 cycle PCR with 454 adapter primers, Lane 9: no-DNA control.

The unexpectedly small apparent size of the 15 cycle PCR products raised the possibility that even though the initial template DNA molecules were mostly > 500 bp in apparent size as confirmed by TA cloning, there may have been present a small population of short primer-multimer products. The presence of small primer-multimers (5/5 clones) was confirmed by TA cloning of these 15 cycle products (Appendix IV, Section B). Therefore, even though TA cloning of the template population revealed only *Ciona* 5'-ends, subsequent amplification with 454 adapter primers produced a large majority of short primer-multimer products. Presumably a low level of primer-multimers in the original template were preferentially amplified and thus in later cycles form a majority of the products and are able to form heteroduplexes with real cDNAs and be recovered from gel regions corresponding to larger apparent size.

Assuming that the sequential amplification (first 5 cycles, then 30 cycles with the original primers, then an additional 15 cycles with the 454 primers) introduced too many cycles, PCR with the 454 adapter primers was tried on the large cDNAs fraction(s) from the early 5 cycle PCR gel filtration step (Figure 15). Now when amplifying the cDNAs for 23, 27 or 30 cycles, the DNA amount increased markedly in all cycles, even during cycles 28-30, and appeared at its true size, with no upward shift in apparent size at higher PCR cycle numbers, suggesting that heteroduplex formation should have been minimal (Figure 15). The products from the 30 cycle PCR were gel extracted (> 500 bp), and analyzed by TA cloning. As expected, all 8 clones tested had large inserts and randomly sequencing 2 of the 8 confirmed presence of real cDNA sequence (Appendix IV, Section C, Rows 1-2).

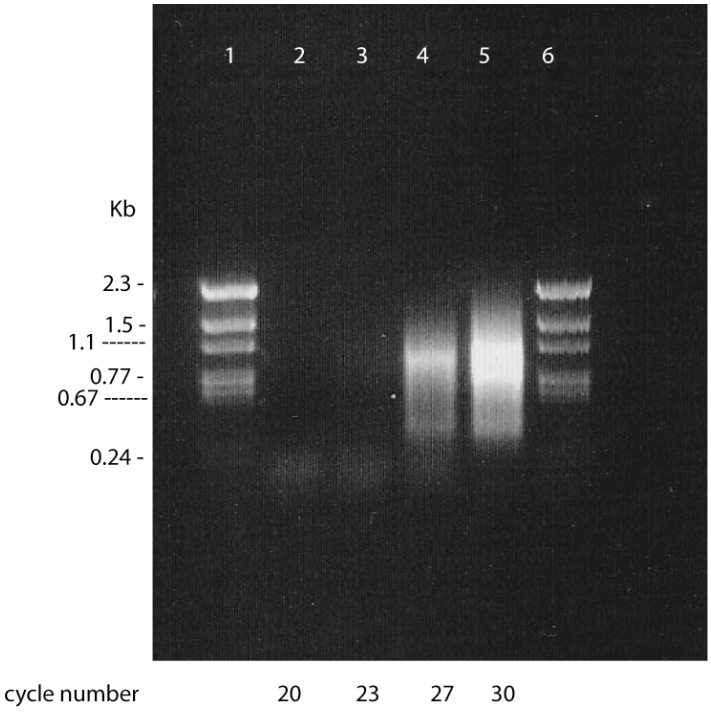


Figure 15: PCR with 454 adapter primers using as template larger *Ciona* oligo-cap 5'-RACE products recovered by gel filtration chromatography of an initial 5 cycle PCR reaction. Product apparent sizes are consistent between lower and higher cycle number and the amount of DNA increases in subsequent cycles. These are signs that amplification is within the exponential phase and has not reached the plateau phase in which heteroduplex formation could occur. Lanes 1 and 6: DNA size ladder, Lane 2: 20 cycle PCR, Lane 3: 23 cycle PCR, Lane 4: 27 cycle PCR, Lane 5: 30 cycle PCR.

The 30 cycle PCR with 454 adapter primers immediately after early-cycle gel filtration was therefore applied to all other libraries (embryo and adult *Petromyzon*, embryo and adult *Branchiostoma*, embryo *Saccoglossus* and embryo *Strongylocentrotus*) and the 5'-ends of all libraries confirmed by TA cloning and sequencing (Appendix IV

Sections C, Rows 1-15). In all cases, except for one, the TA clones corresponded to bona fide 5'-RACE cDNA products.

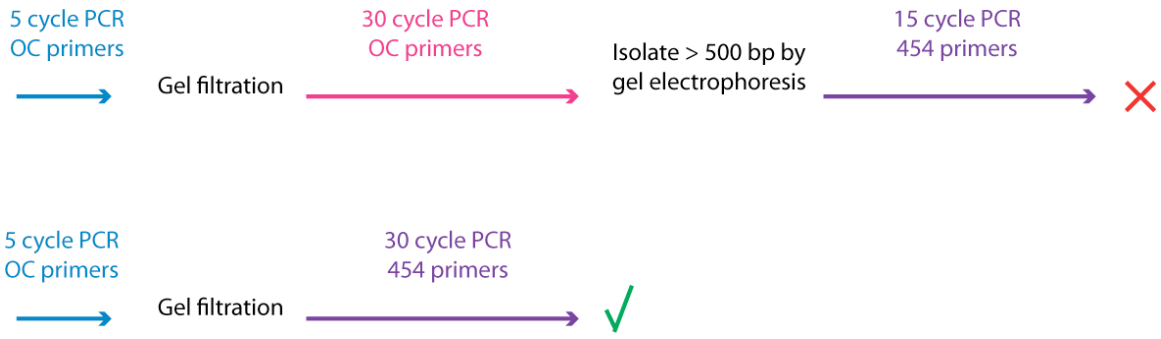


Figure 16: Schematic of PCR steps used to make cDNA libraries with apparent heteroduplexes present (top) and cDNA libraries free of heteroduplexes (bottom). In both cases all DNA molecules TA cloned after 35 total cycles of PCR (with original GeneRacer 5' Nested and reverse primers or 454 primers) were bona fide cDNA clones, green checkmark. However, analysis of 50 cycle products (top) showed the presence of almost exclusively short primer-multimer products, presumably reflecting heteroduplex formation and anomalous gel electrophoretic mobility giving apparent size > 500 bp, red X. OC: oligo-cap primers, GeneRacer 5' Nested forward primer and 3'-anchor-based reverse primer. 454: 454 amplification primers (see Figure 10).

To assess the suitability of our libraries for 454 sequencing, the McGill University Genome Quebec Innovation Centre allowed us to verify our control *Ciona* cDNA population by spiking one of their internal quality control 454 sequencing reactions. Out of 76 reads obtained (See Appendix V), 30 appeared to be real 5'-end cDNAs as assessed by high quality alignments to the NCBI ESTs and the integrated genome, which includes all known SL *trans*-splicing genes (located at http://hoya.zool.kyoto-u.ac.jp/cgi-bin/gbrowse/kh/ , Satou et al., 2005). These 30 reads had a mean read length of 217 bases and represented 13 different genes, of which two had a common 5'-end, the known SL sequence (see Discussion section of this Chapter, Proof of Principle). Only 8/76 consisted of primer-multimers of average size 162 bp, these primer-multimers are believed to have originated from a small number of contaminating heteroduplexes. However, the other 38

reads represented something we had not yet observed, simple primer-dimers. Unlike the primer-multimer products previously discussed, primer-dimer products include only a single copy of the forward primer (and a single copy of the reverse primer). Therefore optimizaton was still required.

Interestingly, these 38 sequences were not classic primer-dimers. The 3'-end of the forward 454 adapter primer ends four bases upstream of the end of the oligo-capping RNA oligonucleotide. A classical PCR primer-dimer is generated strictly by interactions between the primers and should be template independent. However, all of the 454 sequenced primer-dimers contained at least 3 bases, GAA of the GAAA motif, which is the 3'-end of the oligo-capping RNA oligonucleotide, but is not present in the forward, or reverse primer (Figure 17). Thus, the 38 "primer-dimer" molecules arose from amplification of template molecules.

Figure 17: Primer-dimers seen in *Ciona* oligo-cap 5'-RACE cDNA 454 preliminary sequencing data. The schematic shows the general structure of the primer-dimers with example primer-dimer sequence. The GAA motif had varying number of A residues and the 5'-end of the reverse primer (ATATGA…) was often 1-2 nt truncated, as seen in this example. The blue box is the forward primer. The GAA motif originates from the RNA oligoribonucleotide used in the oligo-capping reaction. The green box is the reverse primer.

Such primer-dimers had not been observed in our prior TA cloning and sequencing of the PCR products, perhaps because the cloning procedure interrupts and inactivates a lethal gene permitting positive recombinant selection, so that small inserts, such as primer-dimers, would be less likely to inactivate the lethal gene and thus to survive.

Overview of raw 454 sequencing data

For our main 454 sequencing run all the libraries were pooled based on molar proportions so that each species was equally represented by 25% of the DNA molecules to be sequenced. A previous study in *Ciona* comparing the *trans*-splicing status at different life stages (embryo vs. post-metamorphic juvenile) found that while all stages exhibited the same amount of non-*trans*-spliced and *trans*-spliced genes, there was a higher ratio of *trans*-spliced mRNAs in the embryonic stage than later in development (Sierro et al., 2009). Thus, although unlikely that different developmental times influence the overall SL *trans*-splicing status of a species, where available, a small percentage of adult RNA was also assessed in case other species exhibit dramatically different *trans*-splicing profiles from that of *Ciona*.

For those species with both an adult and embryo library, *Branchiostoma* and *Petromyzon*, emphasis was placed on the embryo libraries, representing 20% each, with the adult libraries representing the last 5%. Additionally, in order to optimize the results, an AMPure Beads (beads coated with positive charge) binding and elution step was included in the hope of removing any "simple" primer-dimers based on their small size having a lower total negative charge than real cDNA molecules. Following emulsion amplification and bead enrichment for DNA-containing beads, a half-plate 454 sequencing run was completed.

The Roche 454 GS-FLX Titanium instrument interprets the raw data in terms of DNA sequence and internal quality control filters. Interpretable sequences that pass the quality control filters are trimmed to remove any terminal nucleotides attributable to the 454 amplification primers and are reported as individual reads. The instrument has two alternative software pipelines for this overall process, the "shotgun" pipeline and the "amplicon" pipeline (Staffa, personal communication). For each library the amplicon

pipeline produced more final reads than did the shotgun pipeline (Table 3). A preliminary analysis of the *Branchiostoma* adult (MID-2 library) sequence data indicated that both pipelines produced some reads not produced by the other (12/22 randomly-selected "shotgun" reads were also present in the larger "amplicon" read set, but 10/22 were unique to the smaller "shotgun" read set). Therefore, in all 5'-RACE library 454 sequence analyses we pooled both read sets into a single combined read set.

Table 3: Numbers of reads for each oligo-cap 5'-RACE library from 454 sequencing

MID Library	Species (stage)	Amplicon pipeline reads	Shotgun pipeline reads	Final filtered reads
MID-10	*Strongylocentrotus* (embryo)	243	0	106
MID-7	*Saccoglossus* (embryo)	98342	48095	298
MID-2	*Branchiostoma* (adult)	308	140	1256
MID-3	*Branchiostoma* (embryo)	2736	575	
MID-4	*Petromyzon* (adult)	124	0	1856
MID-11	*Petromyzon* (embryo)	1817	946	

A half-plate 454 GS-FLX Titanium sequencing run is expected to generate 4-5 x 10^5 reads. Given the intended equimolar mixing of 5'-RACE libraries by species, we expected 10^5 reads for each of the four species investigated. However the instrument reported only ~10^5 total filtered reads (See Table). By sorting the reads according to their MID barcodes it was noted that the great majority were derived from the *Saccoglossus* (MID-7) library. The adult and embryo *Branchiostoma* (MID-2 and MID-3) libraries combined accounted for ~3 x 10^3 reads, while adult and embryo *Petromyzon* (MID-4 and MID-11) libraries combined accounted for ~2 x 10^3 reads. Lastly, only ~2 x 10^2 reads were recorded for embryo *Strongylocentrotus* (MID-10) library. The lower than

anticipated total number of reads and the unequal distribution among the four species suggest that the 454 sequencing approach for these libraries requires optimization (See Discussion).

It was also possible to recover a third instrument output, namely the raw unfiltered reads before any processing by either the amplicon or shotgun pipelines. This unprocessed set consisted of ~ 6 x 10^5 reads and they were of sufficient quality to clearly show that the vast majority were, like the vast majority of the reads that successfully passed through the amplicon and shotgun pipelines, derived from the *Saccoglossus* MID-7 library. It was unexpected that the great majority of DNA molecules analyzed would derive from just one of the libraries analyzed.

The reason the majority (~85%) of raw reads failed the quality control filters in the amplicon and shotgun pipelines is not clear, but many of the raw reads showed oligo-cap primer sequence variants that could be consistent with a failure to fully fill in the nascent DNA chain when two G's in a row were required (results not shown). Thus it is possible that in many picotitre plate wells the library DNA molecules were amplified to such high levels on the bead that there was nucleotide substrate depletion within the picotitre reaction well. Technical aspects of the 454 sequencing are considered below (see Discussion section of this Chapter).

Initial manual analysis of MID-2 reads

Although further optimization is required to maximize sequence data recovery in future analyses, there were nonetheless a significant number of reads that did pass the quality control filters. To assess the suitability of these reads for revealing mRNA 5'-terminal sequences I performed an initial manual analysis of the reads from the

Branchiostoma adult (MID-2) library which comprised 448 reads, 308 from the amplicon pipeline and 140 from the shotgun pipeline.

The shotgun and amplicon reads were combined to form a single list from which all primer-multimer or primer-dimer products identifiable by eye were deleted, thus reducing the number of reads from 448 to 418. These 418 sequences were trimmed of their 5' oligo-cap primer, including all 5' A residues (see Chapter 2) and then each was used as the query in web-based BLAST searches of the NCBI *Branchiostoma* (taxid: 7737) nucleotide collection (nr) and EST databases. For each read the accession number of the top hit, Unigene ID number if reported on the BLAST output, and name of gene if identified in the Unigene record, were recorded. In general, alignments were of excellent quality (see also below for genomic alignments by BLAT), however there were 48 reads for which no significant alignment was found; these were discarded. The remaining 370 reads were examined for duplicates by looking at their gene names, and for each multiply-represented mRNA species a representative example read was chosen (see Methods for details). This resulted in a total of 149 unique mRNA sequences showing that the MID-2 library sequence data represented a significant number of *Branchiostoma* genes.

We noted during the alignment procedure that the MID-2 reads extended further 5' than any aligning EST in the EST collection (an example is shown in Figure 21). This is consistent with the expectation that 5'-RACE products would include the mRNA extreme 5'-terminus, whereas conventional cDNA ESTs would be lacking nucleotides, usually at least 8-21 nt, at the 5'-end (D'Alessio & Gerard, 1988). The extent of 5'-completeness of the MID-2 library products was further assessed by searching for a functional mRNA sequence that, when present, is known to be located at the extreme 5'-terminus, namely the 5'-TOP sequence (Hamilton et al., 2006). The 5'-TOP sequence

functions as a translational control mechanism and is found directly downstream of the cap structure on mRNAs encoding ribosomal proteins and translation factors (Hamilton et al., 2006). The motif begins with a C at the mRNA's 5' terminus followed by 4-14 pyrimidine bases (Hamilton et al., 2006). Translational regulation of TOP mRNAs involves growth factor signaling through the mTOR and PI3K pathways and recognition of the TOP motif by one or more *trans*-acting factors (Hamilton et al., 2006; Orom, Nielsen, & Lund, 2008; Damgaard & Lykke-Andersen, 2011).

I used this 5'-TOP motif as a quality control measure for 5'-completeness of our reads. Of the 370 MID-2 BLAST-alignable reads, 86 reads, corresponding to 32 genes, were identified as being ribosomal protein or translation factor mRNAs (see Appendix VII). Of those reads, over 95% (82/86) had the minimum requirements of a TOP sequence, that is a 5'-end sequence beginning by a C and followed by at least 4 pyrimidines (Hamilton et al., 2006). Therefore, we were able to confirm a very high success rate for our 5'-RACE oligo-capping procedure in terms of identifying the mRNA extreme 5'-terminal sequences. The 4/86 cases of ribosomal protein and translation factor mRNA 5'-RACE products that did not begin with the TOP sequence could represent a low level of unexpected amplification of broken RNAs, as in all cases their 5'-ends aligned to internal positions of their matching NCBI ESTs/cDNAs (results not shown).

This MID-2 manual analysis showed that although the 454 sequencing produced a smaller number of reads than anticipated, the data appeared to be of excellent quality for mRNA 5'-sequence identification and therefore warranted a thorough bioinformatic analysis.

Overview of 454 bioinformatics

In collaboration with Dr. Ken Dewar and Jessica Wasserscheid from the McGill University Genome Quebec Innovation Centre a bioinformatics analysis of the reads from each species was established (See Methods). This process began with non-redundant pooling of the amplicon and shotgun pipelines for each MID library. The forward primers (including those found within reads) and reverse primers, along with their most common variants were then trimmed. Any trimmed read less than 50 nucleotides in length was discarded. The next step was a preliminary alignment of each read to the relevant genome assembly by stand-alone BLAT analysis. From this analysis, each read that did not align from the first nucleotide was inspected for the presence of variant forward primer motifs that had escaped the automatic read-trimming operation. After manual trimming of such motifs, and of any reads now < 50 nt after this trimming, the reads were once again BLAT aligned with the relevant genome. After this definitive BLAT alignment, any read with an identity score of < 55% and < 30 nt aligning to the genome was discarded to create the final filtered read set. The last step in the bioinformatics pipeline was a gene counter function which counted the total number of distinct genes/mRNA species represented by each read set based on non-overlapping genome mapping coordinates.

Interpretation of 454 bioinformatics data

The ultimate goal of this project was to determine if SL *trans*-splicing occurs in any of the analyzed species. For our purposes, a candidate SL is present if mRNAs deriving from at least two different genes have a common sequence at their 5'-end that does not align to the genome at either gene site. In addition, an AG dinucleotide is

expected directly upstream of the putative genomic *trans*-splice acceptor site. The results of this search for SL candidates is presented below, species by species.

***Branchiostoma*:** The *Branchiostoma* read set originated from two libraries, the MID-2 adult library and the MID-3 embryo library. Combined they had over 3700 reads which reduced to 1313 upon informatic and manual removal of primer-related sequences and < 50 nt trimmed reads. In general, the genomic alignments of these reads after informatic and manual primer trimming and length filtering were of very high quality. As shown in Figure 18, 79% (1024/1313) of reads aligned with % identity scores > 90 (% identity is a custom parameter in which the total number of identical nucleotides in the aligned region is divided by the total length of the query trimmed read). Thus most alignments were within 10% of perfect end-to-end matches of the query read with genomic sequences, indicating that the majority of the reads were high quality sequences corresponding to authentic *Branchiostoma* transcripts. However, a small number (57) of the 1313 read alignments had very low % identity scores (< 55%), and also had less than 30 bases aligning to the genome, and were likely not bona fide alignments, but accidental similarities of unrelated sequences (Figure 18). These were also discarded to generate a final filtered read set of 1256 *Branchiostoma* reads. From these 1256 reads, 532 genes were counted.

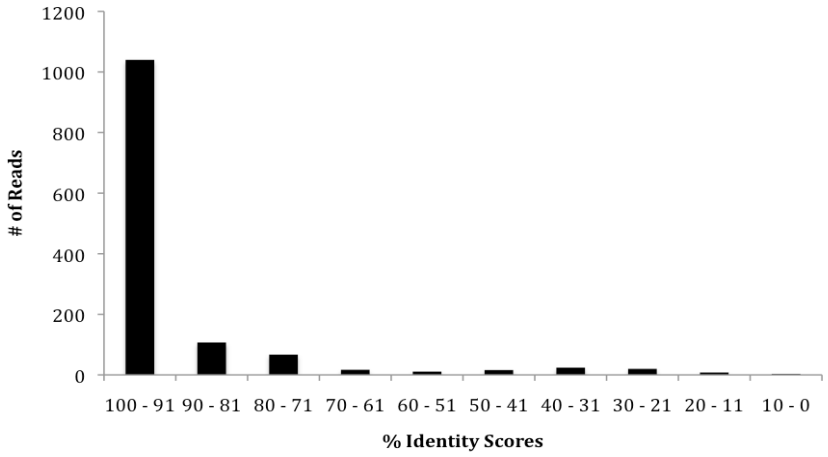


Figure 18: Distribution of percent identity scores during the definitive genomic BLAT alignment of 1313 trimmed and length-filtered (> 50 nt) *Branchiostoma* (MID-2 & MID-3) oligo-cap 5'-RACE reads.

The great majority of reads, 83% (1045/1256), aligned to genomic DNA sequences from the first (or second, or third) base of the read. This is the expectation for conventionally-expressed, non-*trans*-spliced mRNAs. Of the 1256 reads, 220 had 3 or more nucleotides at their 5'-end that were not included in the aligned segment (non-aligned sequences shown in Appendix VIII). Each such non-aligning 5'-end was further investigated. Many (105/220) of these 5'-end alignment failures were due to one or two base mismatches near the 5'-end (an example is shown in Figure 19). Another common cause (20/220) of 5'-end alignment failure was that the 5'-end was in fact present in the genome, but was reported as a separate hit from the top-scoring hit. This could reflect inaccurate genome assembly (see Figure 20). The remaining 95 reads with 5'-ends that failed to align to the genome represent mRNAs whose 5'-segments are either not present in the top 5 hits examined in the BLAT output or are not present in the genome. All reads with non-aligning 5'segments ≥ 3 nt were included in our examination of possible SL candidates.

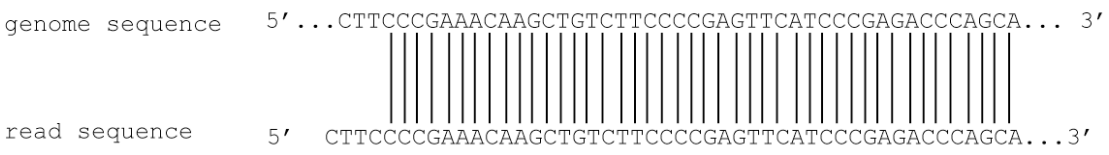


Figure 19: Near 5'-end mismatch causing 5'-end alignment failure. The bottom line shows the 5'-sequence of a *Branchiostoma* 368 base trimmed read. The top line shows the genome sequence aligning with this part of the read (aligning bases indicated by vertical lines). BLAT did not include the first 4 bases of the read in the aligned region, apparently because the read has an extra C residue compared with the genomic sequence.

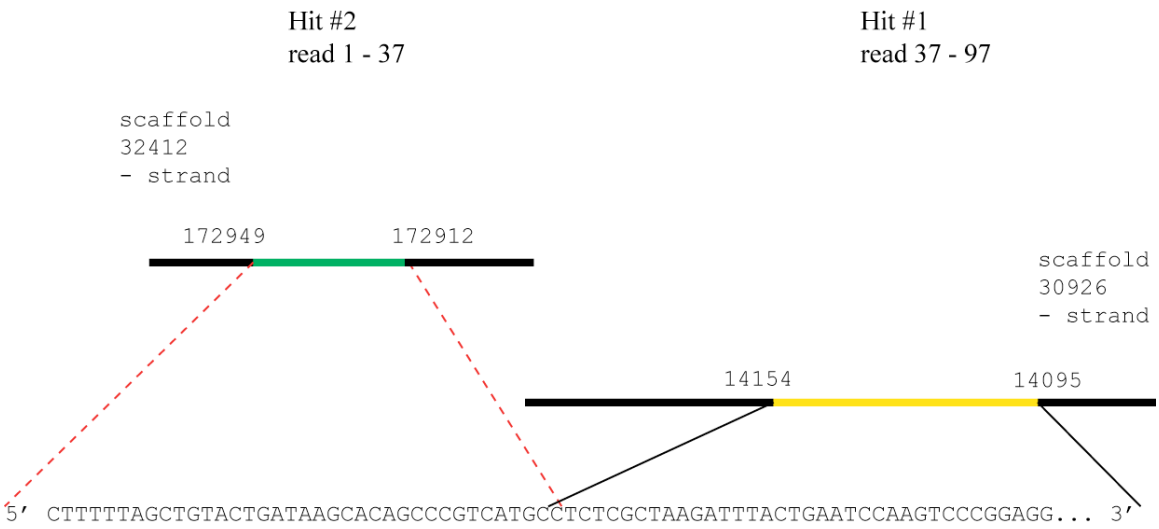


Figure 20: Example of 5'-end of read not aligning with the top hit sequence found in genome. The top hit of a *Saccoglossus* 101 base read (5' sequence on bottom line) to the genome aligns bases 37 – 97 of the read to scaffold 30926 on – strand from position 14154 to 14095. The second best hit aligns the bases 1 – 37 of the 5'-end of the read to scaffold 32412 on – strand from position 172949 to 172912. The presence of a GT splice donor site in the genome at position 172913 and 172914 on scaffold 32412 and an AG acceptor site at position 14156 and 14155 on scaffold 30926 suggests this is an example of mis-assembly of a short first exon. The inclusion of base 37 (C) in both alignments may be due to a missing C base in the read (i.e., CC instead of CCC).

The 220 *Branchiostoma* 5'-end sequences failing to align to the genome were sorted alphabetically as text and examined for a common 5'-end motif which could

represent an SL. There was only one case in which two genes appeared to share a common non-aligning 5'-end. The data suggested that mRNAs derived from two different genes (according to the gene counter, which is based on genomic position) had a common 18 base 5'-end that did not align with the genome, and as such could be considered an SL candidate (Figure 21). However, these two "genes" were possibly one and the same gene. BLAST alignment of one read against the other showed 95% identity starting from the first base of each read and extending to the last base (nucleotide 117) of the shorter read. Using the reads as queries against EST data it was revealed that they both encode for the 60S ribosomal protein L28 isoform 2. Furthermore, the common 18 nt non-aligning 5' sequence includes a 5'-TOP motif at the 5'-end of the reads (see Figure 21), which is expected for a non-*trans*-spliced ribosomal protein gene (*trans*-splicing would be expected to remove the original 5'-end of the pre-mRNAs including any 5'-TOP sequence). The presence of this gene at two genomic sites could be due to erroneous assembly of maternal and paternal alleles as distinct loci at separate locations in the genome, or the occurrence of near-identical gene family members. The absence of the 5'-RACE product 18 nt 5'-end segments from among the top 5 genomic hits could be explained if the 18 bases formed a first exon missing from the genome assembly, or if an 18 nt exon with perhaps one or two mismatching nucleotides could not be identified by BLAT. The 18 nt mRNA 5'-segment is clearly present, in a 5' truncated version, in the *Branchiostoma* conventional EST database (Figure 21). This cannot be considered a strong SL candidate because it is not clearly found on mRNAs deriving from different genes. As this was the only potential SL candidate, it can be concluded that SL *trans*-splicing does not occur at appreciable levels in *Branchistoma floridae* (see Discussion section of this Chapter).

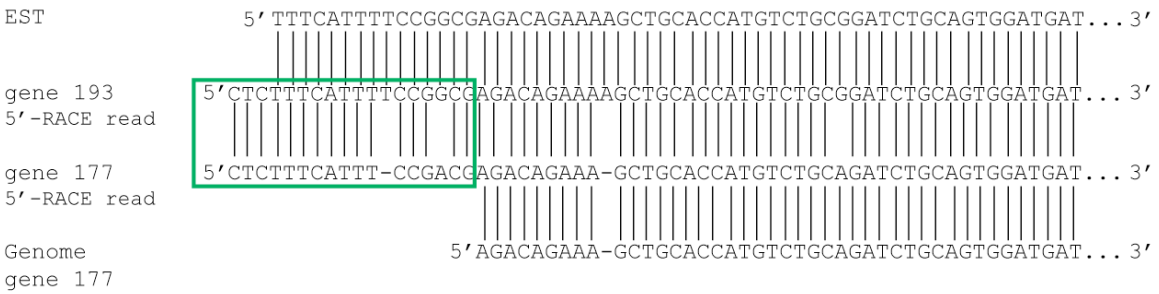


Figure 21: A candidate SL sequence (green box) at the 5'-ends of two *Branchiostoma* oligo-cap 5'-RACE reads likely represent two instances of a single, or nearly identical, non-trans-spliced 5'-TOP mRNA species encoding ribosomal protein L28. The gene counter function classified these two reads as two separate genes, gene 193 and gene 177. When aligned to the genome (bottom line) the first 18 bases (17 bases for gene 177) did not appear in any of the hits. However, when aligned to EST data (top line), the 5'-end is present, although truncated, as expected (because of the 5'-RACE procedure utilized to make these products, it was generally the case, as seen here, that our sequences went further 5' than the 5'-most conventional EST in the NCBI EST database). It is possible that the first 18 bases correspond to a 20 nt long first exon that ends in AG, while the genomic AG could mark the acceptor site for this first exon. This short first exon is either missing from the assembled genome or fails to align because of short length and nucleotide mismatches. BLAST alignment of gene 193 read (369 bases) and gene 177 read (117 bases) reveals 95% identity between the two sequences and both link to the same Unigene ID for the 60S ribosomal protein L28 isoform 2. Note the presence of the 5'-TOP sequence CTCTTTC at the 5'-ends of the 5'-RACE reads.

***Saccoglossus*:** The *Saccoglossus* MID-7 library produced ~10^5 reads, however, once processed by bioinformatic and manual primer trimming and length filtering, the number was reduced to 545 reads. These 545 reads showed a bimodal distribution of genome alignment quality scores; they were very high or very low (Figure 22). Given that the great majority of the initial set of MID-7 reads corresponded to primer-multimers and primer-dimers, it is possible that many of the low quality alignments could represent a minor component of primer-multimers or primer-dimers having unusually divergent sequence variants and thus escaping the filters, including the manual search of non-aligned 5'-ends for motifs resembling the forward primer after the preliminary BLAT alignment. Upon elimination of reads with identity scores < 55% and alignments < 30

bases, the number was reduced to a final filtered read set of 298 high quality alignments, from which 79 genes were calculated using the gene counter.

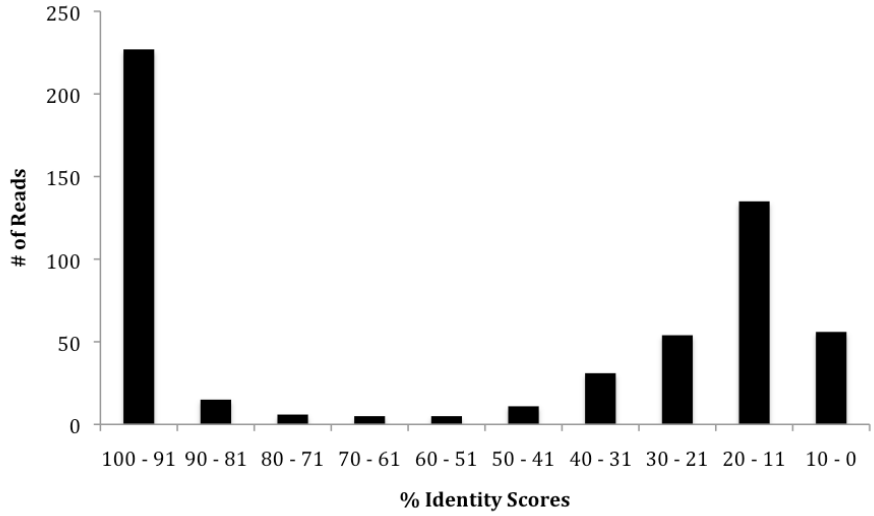


Figure 22: Distribution of percent identity scores during the definitive genomic BLAT alignment of 545 trimmed and length-filteres (> 50 nt) *Saccoglossus* (MID-7) oligo-cap 5'-RACE reads.

The great majority (270/298) of the final set of reads aligned with genomic DNA from the first (or second, or third) base of the trimmed read. Thus the majority of *Saccoglossus* mRNA molecules are not *trans*-spliced. There were 27 reads with 3 or more nucleotides at their 5'-end that did not appear to align to the genome (non-aligning sequences are presented in Appendix VIII). Each non-aligning 5'-end was investigated individually. Fifteen of these non-matching sequences appeared to be due to one or two base mismatches or short first exons not being found in the top hit match.

The 27 non-aligning 5'-ends were sorted alphabetically and visually inspected for a common sequence. There were no common non-aligning 5'-end sequences and hence no SL candidates. Thus *Saccoglossus kowalevskii* does not appear to utilize SL *trans*-splicing (see Discussion section of this Chapter).

***Strongylocentrotus*:** The *Strongylocentrotus* MID-10 library read set consisted of only 243 total reads. Bioinformatic and manual primer trimming and length filtering reduced the set to 117 reads. The great majority aligned to the genome with high % identify scores, while only 9 had identity scores < 55% and alignments < 30 bases and were discarded, thus reducing the number of reads to 106 (Figure 23). From these 106 reads, 71 genes were counted.

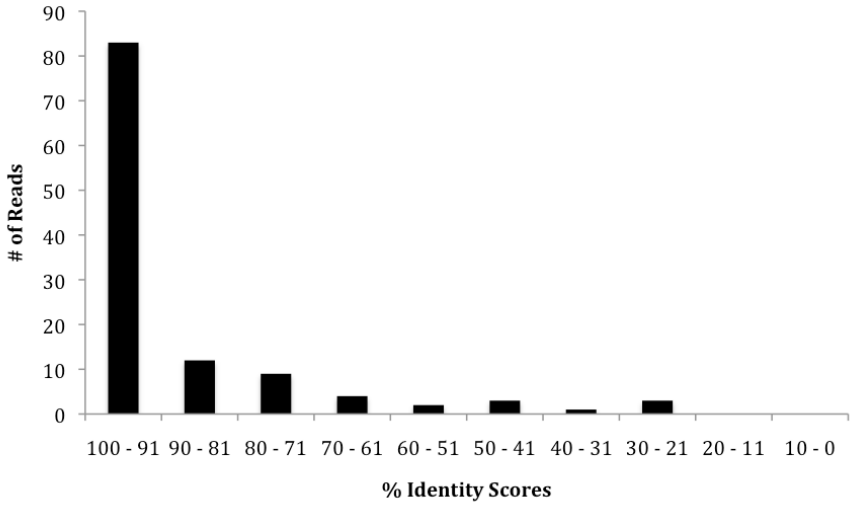


Figure 23: Distribution of percent identity scores during the definitive genomic BLAT alignment of 117 trimmed and length-filtered (> 50 nt) *Strongylocentrotus* (MID-10) oligo-cap 5'-RACE reads.

Of the 106 reads, 34 had 3 or more nucleotides at their 5'-end that did not appear to align to the genome (non-aligning sequences are presented in Appendix VIII). Twelve of the non-matching sequences were due to one or two base mismatches or short first exons present in the genome but not being associated with the top score hit. There were no common non-aligning 5'-end sequences and hence no SL candidates. Though the number of reads is modest, this data suggests that *Strongylocentrotus purpuratus* does not utilize SL *trans*-splicing (see Discussion section of this Chapter).

***Petromyzon*:** The MID-4 adult and MID-11 embryo *Petromyzon* libraries had a combined total of over 2880 reads. Following bioinformatic and manual primer trimming and length filtering the number was reduced to 1974. Of these, 119 had very low % identity scores and less than 30 bases aligning to the genome and thus were discarded to make the final filtered read set of 1856 reads (Figure 24).

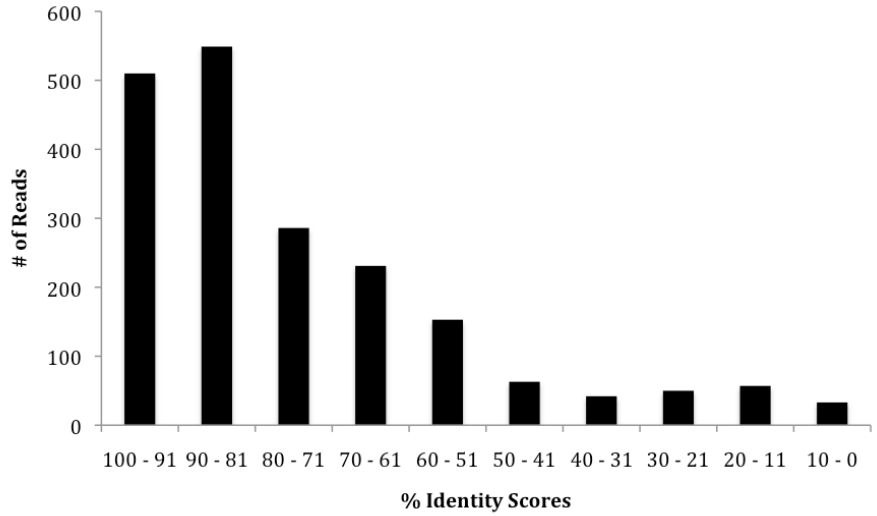


Figure 24: Distribution of percent identity scores during the definitive genomic BLAT alignment of 1974 trimmed and length-filtered (> 50 nt) *Petromyzon* (MID-4 & MID-11) oligo-cap 5'-RACE reads.

Puzzlingly, the great majority of these 1856 reads with strong genomic alignments did not appear to be typical gene/mRNA sequences, but were composed of simple dinucleotide or trinucleotide repeats (Figure 25A). Because they align more or less well to the genome, they could not be easily distinguished or separated by our methods from authentic cDNAs, which presents an impediment to a complete analysis of authentic 5'-RACE products. A small fraction of the *Petromyzon* reads did appear to correspond to typical mRNA squences; we found 20 – 40 reads with a clear 5' TOP sequence, including reads that mapped to the known genes in the genome (Figure 25B). Because our bioinformatic approach does not separate these from the simple repeat sequences, it will

require extensive manual analysis to recover and study only the set of authentic mRNA 5'-RACE products.

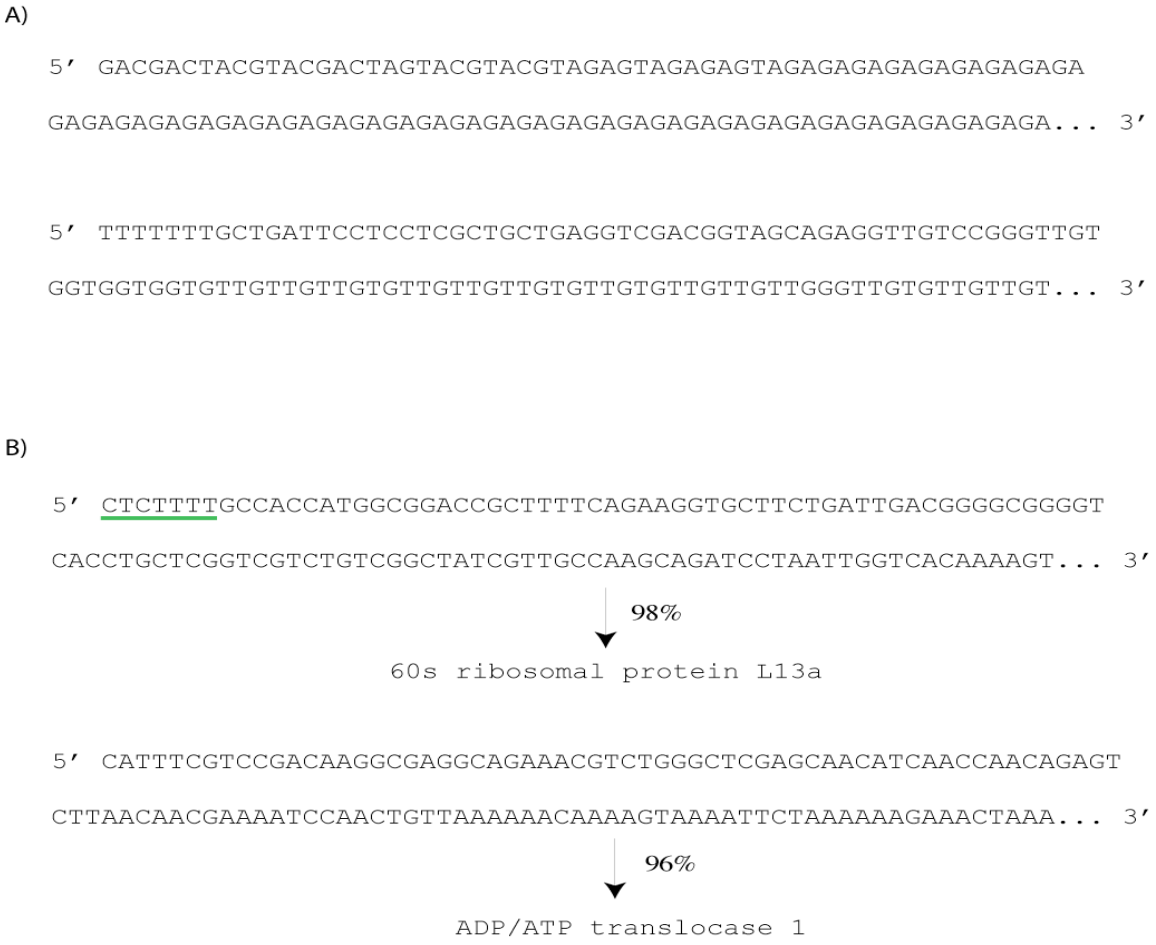


Figure 25: Examples of trimmed *Petromyzon* 454 sequencing reads. The first 120 nt of each read are shown. A) A significant fraction of the reads resembled these two examples of simple dinucleotide or trinucleotide repeat reads. B) The *Petromyzon* reads also included some authentic cDNA molecules that aligned to known genes (found by manually aligning multiple randomly chosen reads). Two examples are shown with their corresponding gene name, and NCBI BLAST identities scores, as determined by EST alignment and associated Unigene links. Green underline indicates 5' TOP sequence.

DISCUSSION

PCR and primer-multimer products

Our analysis of the *Ciona* oligo-capped 5'-RACE library, made as a control for the ability of the procedure to detect *trans*-splicing where it is known to occur, revealed an interesting problem. Although we gel-isolated large PCR products (300 – 700 bp)

when we analyzed these by TA cloning and Sanger sequencing we found the majority (63/74) to be small (~150 bp) primer-multimers. The primer-multimers were likely made during the oligo-capping procedure by RNA ligation of 5'-truncated oligo-cap RNA oligonucleotide molecules.

The unexpected short primer-multimer products present in the gel-isolated 300 - 700 bp 5'-RACE PCR products could have two sources. Firstly, as these small primer-multimers migrate more quickly than the authentic 5'-RACE products during agarose gel electrophoresis it is possible that some molecules may get snagged in the gel and thus be recovered, presumably as a very small minority component, in the population of isolated large products.

The second possibility is that the primer-multimer products were migrating anomalously slowly during gel electrophoresis, perhaps as a result of heteroduplex formation. Even though the population of mRNA molecules amplified was likely very diverse, the 5'-ends (the oligo-cap adaptor/primer) and 3'-ends (the reverse primer) were identical in all products. We hypothesized that heteroduplexes based on these identical end sequences were being formed during amplification, in which one strand of the heteroduplex is a primer-multimer while the other strand is a real (and longer) 5' oligo-capped cDNA molecule.

During the plateau phase of PCR which occurs at high cycle numbers, the complementary strands of products made in earlier cycles anneal with each other rather than with free primers, thus contributing to the lack of net DNA synthesis during the plateau phase (Thompson, Marcelino, & Polz, 2002). Likewise heteroduplexes can be formed during the plateau phase by re-annealing between partially complementary products made in earlier cycles (Thompson et al., 2002). When amplifiying multiple DNA molecules with similar sequences, such as alternatively spliced isoform mRNAs,

heteroduplexes form readily as the very similar sequences can easily anneal to each other (Eckhart et al., 1999). The heteroduplexes of two alternatively spliced vascular endothelial growth factor isoforms studied by Eckhart and colleagues appeared to migrate either at an intermediate size (600 bp) or, at very high cycle numbers, much larger size (1200 bp), due to weak interactions between heteroduplexes, than the expected homoduplexes (526 bp and 658 bp). Extrapolating from their results, heteroduplexes formed during amplification of a complex population of 5'-RACE products and primer-multimer products may not run at the true size of either their shorter or longer strand. Slowly migrating heteroduplexes could perhaps account for the presence of short primer-multimer DNA strands in regions of the gel corresponding to perfect duplex DNA of larger size.

Elimination of primer-multimers

We took two steps to eliminate, or at least reduce, the presence of short DNA strands among apparently larger-sized DNA molecules. First, we incorporated a column chromatography gel filtration step to separate 5'-RACE products by size. Unlike gel electrophoresis, where small molecules run faster than large molecules and thus can "contaminate" the agarose from which larger sized molecules will later be isolated, small molecules run slower than large molecules during gel filtration chromatography, so such contamination should not arise. Secondly, we attempted to restrict all PCR amplifications to low cycle numbers, i.e., to the exponential phase of perfect duplex synthesis, thus avoiding the plateau phase during which heteroduplexes could be formed. These two steps of eliminating small molecules (including primer-multimers) and heteroduplex formation were applied to all the oligo-capping 5'-RACE cDNA libraries. TA cloning confirmed that partically all inserts examined (32 representing all 6 libraries) were at

least 500 bp and sequencing of a few examples from each library appeared to confirm the success of these measures to avoid contaminating small products as the isolated large products were real cDNA molecules with no short cDNAs or primer-multimers found (15/15).

However, when these same cDNA libraries were sequenced by the 454 high-throughput method, primer-multimers were a major component of the reads obtained. Primer-multimers were approximately 30% of *Petromyzon* reads, 50% of *Strongylocentrotus* reads, 65% of *Branchiostoma* reads, and 99.5% of *Saccoglossus* reads, even though none were observed when analyzing the libraries by TA cloning.

Therefore, either the TA cloning method under-represented the presence of primer-multimers, or, the 454 sequencing over-represented the number of primer-multimer molecules in the libraries.

TA cloning may have skewed the average molecular size by selectively cloning the larger cDNA molecules. The cloning procedure interrupts and inactivates a lethal gene, permitting positive recombinant selection (Bernard et al., 1994), and perhaps small inserts, such as primer-multimers, are less likely to inactivate the lethal gene and thus less likely to be cloned.

The 454 sequencing method could also introduce a bias, but for smaller, rather than larger molecules. Observing all libraries, the longest reads were ~600 bp. If the cDNA libraries were composed of molecules ranging from 500 – 1500 bp, as intended, then reads in which the sequencing ended before reaching the 3’ primer at the far end would be expected to be commonly observed. But examples of such longer 5’-RACE cDNA molecules were practically non-existent in the six libraries as almost all reads examined contained the 3’ primer. Thus, almost no 5’-RACE products that were successfully amplified during the bead-coupled emulsion PCR amplification step of the

454 method were longer than 600 bp. It therefore appears that most of the products in our 500 – 1500 bp cDNA library fractions were simply too long to be amplified during emulsion PCR. If this is the case, then most beads with cDNA would fail to amplify the molecule, and they would consequently be eliminated in the bead enrichment step which is designed to discard beads without DNA. Meanwhile, a perhaps smaller number of beads with short molecules, such as primer-multimer products would amplify well and thus be over-represented after the bead enrichment step.

Of course, although presented as contrasting options, it is possible that the real nature of these cDNA libraries lies somewhere in between the two extremes. Perhaps the TA cloning does under-represent the number of primer-multimers in the libraries, but so too does 454 sequencing over-represent their presence in the libraries.

Future oligo-cap 5'-RACE libraries

Optimization of the production and sequencing of these libraries to produce a large number of real 5'-end cDNA reads remains a major goal for further pursuit of high-throughput 5'-RACE sequencing. As discussed above, it is possible that excessive 5'-RACE product length exceeding the potential of emulsion PCR amplification is a major contributing factor to the over-representation of short primer-multimers. If this is the case, then increasing the emulsion PCR extension time which could permit amplification of longer molecules may improve results. Furthermore, in conjunction with increased emulsion PCR extension time, isolating DNA up to 1000 bp, instead of 1500 bp as done in this set of experiments, may also be considered in order to reduce the proportion of molecules too long to amplify. Decreasing the lower limit to below 500 bp is not advisable, as we believe on the basis of TA cloning data for these products that a significant proportion of < 500 bp molecules are primer related sequences.

Another way of perhaps avoiding the failure of the emulsion PCR amplification of longer products, and thus reducing the dominance of primer-multimers, is by completing the oligo-capping, then 5 cycle PCR and column chromatography as previously done, but then, instead of amplifying the products, shear the DNA to a smaller average size, e.g. 250 bp and ligate the 454 adaptors directly to the sheared DNA fragments. In this case, many reads will originate from sheared internal fragments and not include mRNA 5'-terminal sequences, but the mRNA 5'-ends will be identified as those reads beginning with the oligo-cap adapter sequence. In this method, although only a fraction of reads would reveal mRNA 5'-end sequences both long and short 5'-RACE products would be well represented in the reads, so that an initially small component of primer-multimer or primer-dimer products could not dominate the results.

In addition to the 454 high-throughput sequencing results revealing primer-multimers to be a much more prevalent problem than anticipated, the results also suggested that there are additional points where further optimization is required. More than 3×10^5 picolitre sequencing reaction wells failed to pass the shotgun and amplicon pipeline quality filters. These failed reactions may be due to an excess amount of DNA present on many beads. If as predicted, each reaction well contains one bead, but on that bead there is an excessive amount of DNA, then it is possible that the PCR components, such as nucleotides, gets depleted within that well. Thus, if there is only enough of the correct nucleotide for half of the DNA molecules on the bead to be correctly sequenced than the other half will fail to incorporate a nucleotide for that flow cycle. This will cause the next round of nucleotide incorporation to be one base behind for half of the molecules and as sequencing rounds progress every round will incorporate at least two nucleotides, (even though the bead only contains multiple copies of the same DNA molecule), thus the sequencing will consequently be considered a failure. Thus, the amount of DNA per

bead should be reduced in any subsequent experiments. This could be accomplished by reducing the number of emulsion PCR cycles.

Interestingly, of the reads that did not fail, the vast majority, ~95%, were from the *Saccoglossus*, MID-7 library. This suggests that the initial pooled library mixture was actually much more heavily composed of MID-7 cDNA than the equimolar intended mixing based on the 5'-RACE products apparent sizes estimated by agarose gel electrophoresis. Perhaps the MID-7 library had a much higher fraction of small primer-multimer products hidden as heteroduplexes. Alternatively, and less likely, the libraries could all have similar proportions of occult small primer-multimers, but only MID-7 ones amplify well, due to some effect of the MID barcode sequence which represents the only difference between primer-multimer products in the various libraries.

Simple repeat *Petromyzon* reads

The highly repetitive sequences that dominate the two *Petromyzon* 5'-RACE libraries will require further investigation. Such simple repeat reads were not seen in any of the libraries for the other species. In *Petromyzon* sequences obtained from TA cloning and Sanger sequencing, these simple repeated DNAs are present, albeit at a much lower level (4/20). Furthermore, closely related sequences are found in the *Petromyzon* genome as well as in ESTs (results not shown). The reason for which these simple repeat 5'-RACE products so strongly dominate the 454 sequencing output remains to be determined. It would be of great interest to further investigate these sequences and determine if they have a real function in *Petromyzon*.

In addition to reads representing simple repeat sequences there is a minority of authentic 5'-RACE products of protein coding genes among the *Petromyzon* reads. However, it will require laborious manual inspection to separate these from the simple

sequence repeat reads because the latter, as well as the former, show significant alignment with the genome.

Proof of principle

From the 454 sequencing experiment in which the McGill University Genome Quebec Innovation Centre spiked one of their internal quality control 454 sequencing reactions with our oligo-capped *Ciona* library we confirmed that our library preparation method would detect SL *trans*-splicing. The experiment provided only 13 different genes, and yet, SL trans-splicing was clearly detected as two of the genes, an uncharacterized locus and zinc finger protein (TRAF/RING)-2 had a common non-aligning 5'-end, corresponding to the known SL. Importantly though, even had the sequence of the SL not been known, or even the *trans*-splicing status of *Ciona*, our data still would have found the SL candidate. Although the sample is small, the ~15% (2/13) of genes found with the SL sequence at their 5'-end appears less than the 50% established from studies of tailbud embryo RNA (Satou et al., 2006). It may be that a smaller fraction of the genes expressed in adult muscle are *trans*-spliced.

Evolution of SL *trans*-splicing in the deuterostomes

Technical issues complicated the 5'-RACE high-throughput sequence analysis. Some species were represented by a relatively small number of reads. However, it was fortunate that *Branchiostoma* was represented by a significant number of authentic 5'-RACE reads. The phylogenetic position of the basal chordate, *Branchiostoma*, having diverged from the ancestral chordate stock before the divergence of the tunicates and vertebrates makes it an especially significant species in which to examine the evolution of SL *trans*-splicing within the deuterostomes. If the trait is ancestral within the

deuterostomes, or within the chordates, then *Branchiostoma* should be an SL *trans*-splicing species. However, if *trans*-splicing arose *de novo* in the tunicates, then the earlier-diverged *Branchiostoma* would not be expected to utilize *trans*-splicing.

The probability of a false negative, i.e., having erroneously missed the occurrence of SL *trans*-splicing in our *Branchiostoma* 5'-RACE products is very low. Certainly, if SL *trans*-splicing were occurring anywhere near the levels observed in *Ciona* embryos (50% of genes generate *trans*-spliced mRNAs [Satou et al., 2006]) it would have been grossly evident in a sample of > 500 genes. The tunicate Oikopleura, has the lowest estimated rate of SL *trans*-splicing with only 25% of genes being SL *trans*-spliced (Ganot et al., 2004). If *Branchiostoma* utilizes SL *trans*-splicing at an even lower level, say for example 10%, the probability of us not finding two or more mRNAs with the SL sequence as a common 5'-end in our sample of 532 randomly selected genes is $p \leq 1.2 \times 10^{-10}$ (see Chapter 2 for the formula for this calculation). Furthermore, even if it were happening on only 1% of the genes, the probability of not observing SL *trans*-splicing in our sample is $p \leq 0.03$. Therefore, we can confidently exclude the possibility of SL *trans*-splicing occurring in the basal chordate *Branchiostoma*, even at very low levels. This is the most important single finding of this thesis research.

Although only 79 and 71 genes were sequenced for *Saccoglossus* and *Strongylocentrotus*, respectively, we can conclude that SL *trans*-splicing is probably not occurring in either species. Assuming SL *trans*-splicing of 10% of the genes, the probability of us not finding two or more mRNA species with a common 5'-end is $p \leq 0.0024$ for *Saccoglossus* and $p \leq 0.005$ for *Strongylocentrotus*. Thus, we can confidently exclude the possibility of SL *trans*-splicing at moderate levels in these species, although

it remains possible that a very small fraction of genes (approximately ≤1%) could be *trans*-spliced.

Collectively these results strongly support the invention of SL *trans*-splicing in the tunicates as no other group deuterostome appears to possess the trait.

CHAPTER 4 – EXPRESSION OF OUTRON-RETAINING CiTnI CONSTRUCTS

INTRODUCTION

In order to examine the 5'-UTR sanitization hypothesis of SL *trans*-splicing, and determine if the outron contains deleterious elements and for this reason is replaced with the SL sequence, an outron-retaining mature mRNA molecule must be experimentally generated and characterized. Our laboratory has characterized and studied in depth the *trans*-spliced *Ciona intestinalis* gene (CiTnI) encoding the muscle contractile regulatory protein troponin I (MacLean, Meedel, & Hastings, 1997; Vandenberghe, Meedel, & Hastings, 2001; Cleto et al., 2003; Khare et al., 2011). In the CiTnI gene the length of the outron is 459 nt (Khare et al., 2011). Previous attempts to make a non-*trans*-spliced, outron-retaining CiTnI mRNA by mutating the natural AG dinucleotide acceptor site (-64) activated cryptic acceptor sites at two nearby positions -76 and -39 (Mortimer, 2007). Complete deletion of the natural acceptor site and putative branchpoint sequence did eliminate *trans*-splicing in the vicinity of the natural acceptor site, but this activated/revealed an additional site of SL *trans*-splicing further upstream (-346) (Mortimer, 2007). The mRNAs being generated thus retained segments of the outron, although the 5'-most part of the outron was in every case removed by *trans*-splicing at the –346 upstream cryptic acceptor site.

In the prior study by S. Mortimer, three constructs were made, each of which produced a transcript that retained a different length of the outron sequence. Results of ß-galactosidase reporter gene activity showed that the mutant with the longest retained outron (-346 to -79, here renamed ABC, see Figure 26) appeared less active than the other two constructs, here renamed AB and A (Mortimer, 2007). I replicated and further advanced these initial experiments by 1) including a transfection control plasmid

construct and 2) by more thoroughly investigating the retained outron, specifically by addressing the question of whether the longest retained outron contained a particular deleterious sequence localized in region C that was present in ABC and absent in AB (specific element hypothesis), or whether it was simply the case that longer retained outrons have a negative impact on gene expression even in the absence of specific deleterious sequence elements (non-specific length hypothesis). First, in order to confirm that S. Mortimer's initial results are indeed due to differences in the constructs themselves and not between electroporations an internal transfection control construct, CiTnI(-1.5 kb)eGFP, was included in the experiment. Then, as the previous results suggest that the outron may indeed contain deleterious elements to the mRNA located in the region -177 to -79 (region C) a fourth construct in the set was engineered. This construct, AC, was designed to contain the same length of retained outron as AB but to contain region C in place of region B and thus reveal whether there are deleterious elements specifically localized in outron region C or if the length of the outron *per se* appears to be the deleterious feature that compromised expression of ABC.

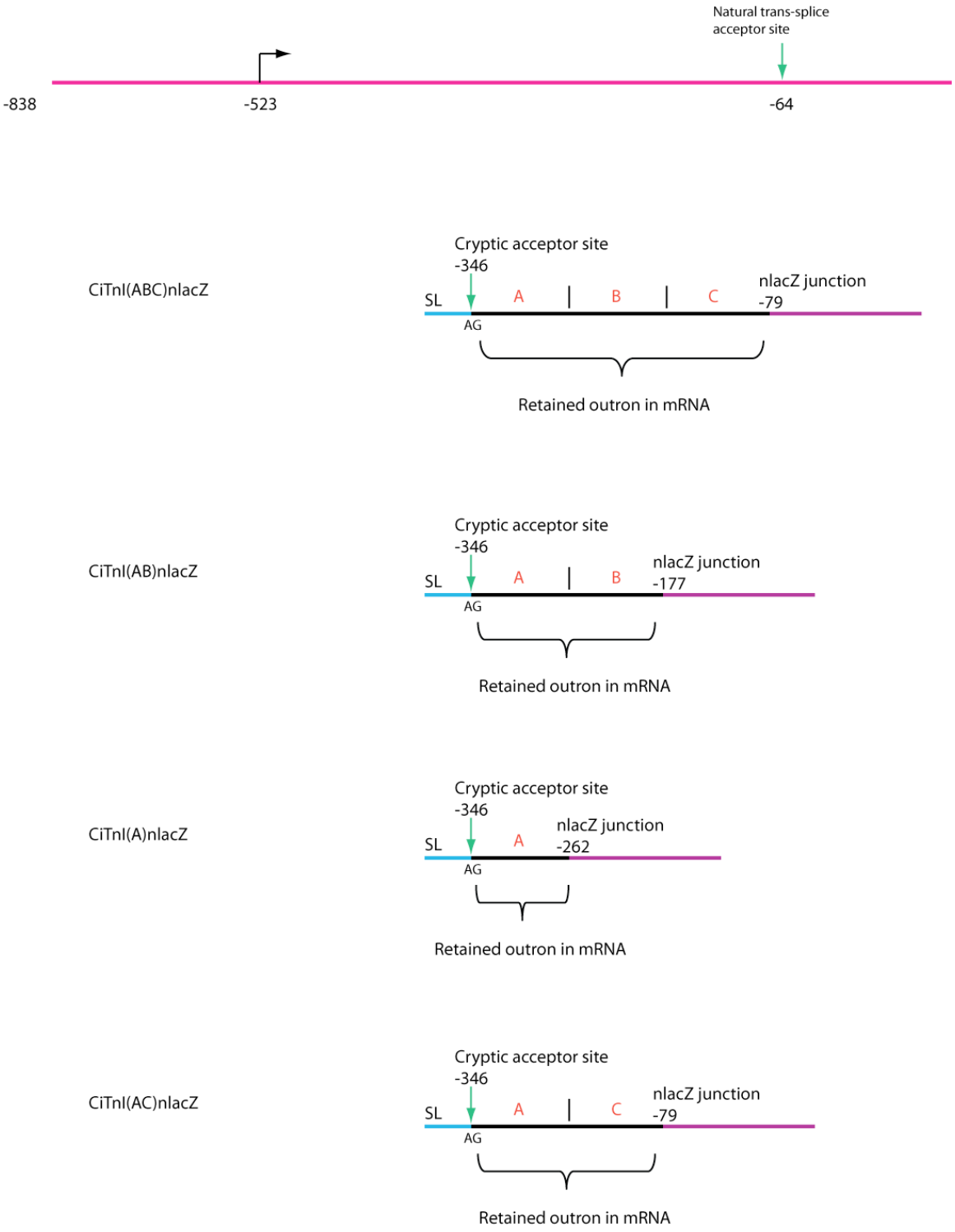


Figure 26: Series of constructs producing different outron-retaining mRNAs. The CiTnI gene is represented by the top line, numbers are in relation to the translation start codon. The sequence from -838 to -523 contains all the DNA elements needed for expression in *Ciona* embryonic muscle cells (Khare et al., 2011). The transcription start site (TSS) is located at -523 and the natural SL acceptor site at -64. The mRNA made by the constructs utilized are depicted below. The first three constructs shown were produced by S. Mortimer (2007), and CiTnI(AC)nlacZ was designed and produced by me for the present study.

RESULTS

Internal transfection control construct validation of S. Mortimer's previous results

In order to validate the results suggesting the construct that retained the longest outron segment, CiTnI(ABC)nlacZ showed a lower level of β-galactosidase activity than those with shorter outrons [CiTnI(AB)nlacZ and CiTnI(A)nlacZ] (Mortimer, 2007) a transfection control construct was needed. I made a GFP reporter driven by the TnI gene regulatory elements, and found that it was expressed as expected in tail muscle cells (Figure 27).

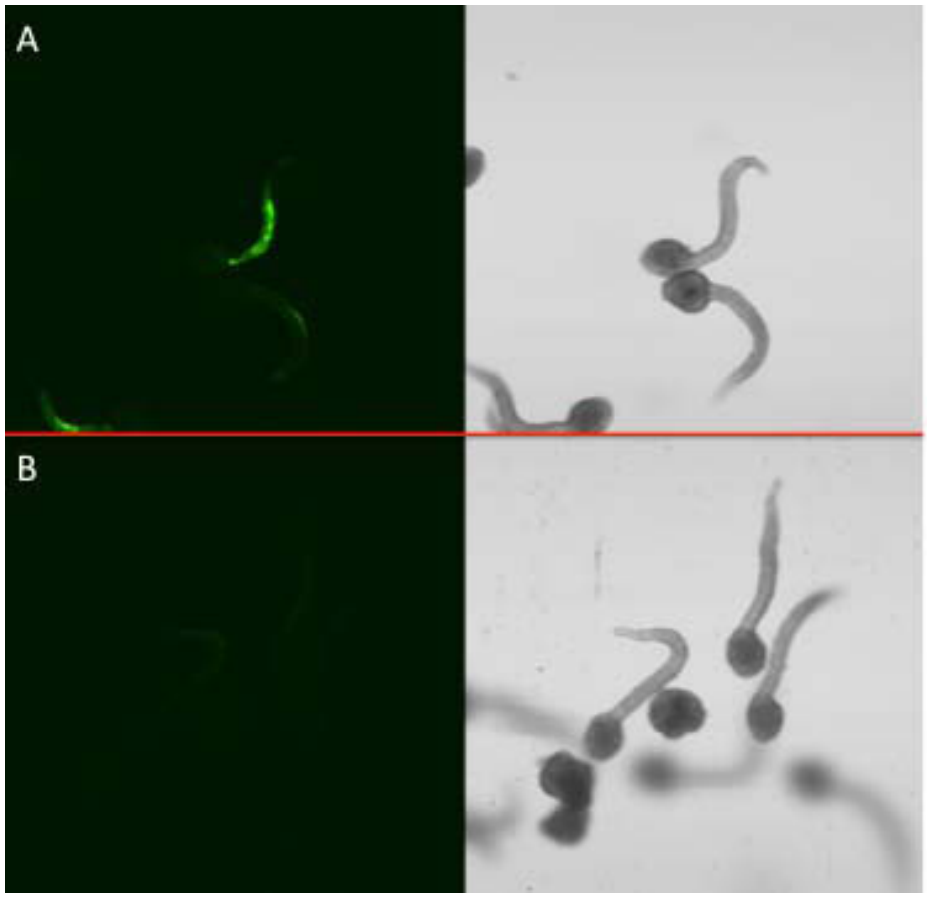


Figure 27: CiTnI(-1.5 kb)eGFP is expressed in *Ciona* embryos. *Ciona* embryos allowed to develop for 12 hours (and stored at 4°C for 6 hours prior to microscopy) were examined under fluorescence (left) and brightfield illumination (right) in an inverted laser-scanning confocal microscope. A) Embryos co-transformed with 12.5 µg of CiTnI(-1.5 kb)eGFP show heterogenous tail muscle cell GFP expression. B) Untransformed embryos show only a low level of autofluorescence.

After validating the GFP construct, I repeated S. Mortimer's experiment but now including 2.5 μg of the CiTnI(-1.5 kb)eGFP transfection control construct in each electroporation was repeated. Three electroporations, one for each CiTnInlacZ reporter construct, were done on aliquots of a single population of fertilized eggs. Embryos developing from the transformed zygotes were examined at 12 h and scored positive or negative for GFP fluorescence using an upright epifluorescence microscope. By this measure, all three groups of transformed embryos showed the same electroporation success rate of 60-70%, as revealed by the number of GFP positive embryos (Table 4). After scoring eGFP expression, the same embryo populations were fixed and stained with X-gal for 2.5 hours to reveal expression of the co-transfected β-galactosidase-encoding outron test constructs. Results of the stain confirmed that even with equal transformation efficiencies as revealed by the GFP reporter, embryos co-electroporated with shorter-retaining outron constructs had much higher reporter gene expression levels than those co-electroporated with the longer-retaining outron construct. The A and AB constructs had 73% (11/15) and 88% (15/17) of embryos in the highest expression category (+++) respectively, while the longer-retaining construct, ABC only had 43% (16/37) of embryos in the highest expression group with 34% (4/37 and 10/37) of embryos showing very little (+), or no staining (0) [Batch 1 in Table 4 and Figure 28].

In subsequent experiments we compared CiTnI(ABC)nlacZ to CiTnI(AB)nlacZ four additional times (Batches 2, 3 and two ABC replicates in Batch 4, Table 4 and Figure 28). With one exception these experimental replicates confirmed that CiTnI(ABC)nlacZ was expressed at lower levels than CiTnI(AB)nlacZ.

To simplify presentation of the results we generated a single-number measure of expression strength, the high/low expression ratio, in which the numerator is the number

(or percentage) of embryos in the strongest expression category (+++), and the denominator is the number (or percentage) of embryos in the weakest expression category (0).

Table 5 summarizes my results and those of S. Mortimer in terms of the high/low expression ratio. In S. Mortimer's initial data, and in my experiments with egg batches 1, 2, and 3, CiTnI(AB)nlacZ gave higher values of the high/low ratio (ranging from 5.4 to >88) than did CiTnI(ABC)nlacZ (ranging from 0.4 to 2.47), indicating higher levels of AB β-galactosidase expression than ABC β-galactosidase expression. With this parameter the difference in expression strength between CiTnI(AB)nlacZ and CiTnI(ABC)nlacZ showed considerable variation, such that AB/ABC relative values within individual experiments ranged from ~2 to >55.

Even greater variation was seen in the electroporations done with egg batch 4. In this batch, the high/low expression ratio for AB was 2.2, the lowest observed in any experiment. This unusually low expression did not reflect a poor transfection because, as measured by co-transfected eGFP, the transfection was very efficient (85%). Moreover, the high/low expression ratio for two replicate electroporations of ABC were 13.4 (the highest observed in any experiment) and 1.6 (similar to the high/low ratio for ABC in the Batch 1, 2, and 3 experiments). This very high expression activity for ABC in one replicate did not appear to be based on unusual transfection efficiency as the eGFP transfection control was also normal. Thus, barring unrecognized experimenter error, there is some variability in gene expression levels in different transfections that is not due to gross differences in transformation efficiency, but to some other unknown biological factor.

However, apart from this one exceptional, 13.4 value, CiTnI(ABC)nlacZ was expressed at lower levels than CiTnI(AB)nlacZ in the remaining 5 of 6 within-experiment

comparisons that can be made in Table 5. Moreover, in two of these cases CiTnI(ABC)nlacZ was expressed at apparently much lower levels than CiTnI(AB)nlacZ. This data confirms S. Mortimer's original conclusion that CiTnI(ABC)nlacZ is more weakly expressed than CiTnI(AB)nlacZ.

CiTnI(ABC)nlacZ, a new outron-retaining construct

Having confirmed that CiTnI(ABC)nlacZ exhibits the weakest β-galactosidase activity, we then included a fourth construct in the set. This construct, CiTnI(AC)nlacZ was engineered to produce the same length of retained outron as the CiTnI(AB)nlacZ construct, therefore testing for any deleterious elements in DNA region C. Four repeats of the electroporations were done and the results of the β-galactosidase activity are summarized in Figure 28 and Table 4 and 5. Briefly, it appears that embryos transformed with the CiTnI(AC)nlacZ, like those transformed with the CiTnI(AB)nlacZ construct, express high levels of β-galactosidase, when compared to the long-retaining outron construct CiTnI(ABC)nlacZ. Furthermore, the high/low ratio values for AC ranged from 22.5 to >98 while those for ABC, even including the unusually high replicate discussed above, ranged from 0.41 to 13.6, clearly demonstrating that AC was always more actively expressed than ABC.

Varying the X-gal staining reaction time

We were concerned that the 2.5 h X-gal stain reaction timemight saturate staining of the β-galactosidase expressing *Ciona* tail muscle cells thus potentially masking real differences in construct expression levels. In order to assess if a shorter staining reaction would be more sensitive in revealing differences in β-galactosidase activity levels, some

embryos were stained for 10 minutes (the time at which the blue stain is just appearing when observing the reaction under a dissecting microscope) not 2.5 hours. Although the shorter staining reveals a lighter blue stain (Figure 29), it does not appear to significantly change the scoring based on counting the numbers of stained tail muscle cells in each embryo, or interpretation of the results (Tables 4, 5 and Figures 28, 29).

Table 4: β-galactosidase expression levels of outron-retaining CiTnI nlacZ constructs in *Ciona* 12h tailbud embryos

Batch of fertilized eggs	DNA Construct	Total number of embryos subjected to X-gal stain	β-galactosidase expression levels				GFP transfection control[A]
			Number of (+++) embryos	Number of (++) embryos	Number of (+) embryos	Number of (0) embryos	
1	CiTnI(ABC)nlacZ*	37	16 (43%)	7 (19%)	4 (11%)	10 (27%)	27/43 (63%)
	CiTnI(AB)nlacZ*	17	15 (88%)	1 (6%)	1 (6%)	0	13/17 (77%)
	CiTnI(AC)nlacZ*	99	90 (90%)	6 (6%)	3 (3%)	0	61/95 (64%)
	CiTnI(A)nlacZ	15	11 (73%)	0	1 (6%)	3 (20%)	15/24 (63%)
2	CiTnI(ABC)nlacZ	24	10 (42%)	7 (30%)	3 (13%)	4 (17%)	-
	CiTnI(AB)nlacZ	19	11 (60%)	4 (21%)	2 (11%)	2 (11%)	-
3	CiTnI(ABC)nlacZ	17	5 (29%)	0	0	12 (71%)	-
	CiTnI(AB)nlacZ	16	13 (81%)	2 (13%)	0	1 (6%)	-
	CiTnI(AC)nlacZ	70	63 (90%)	3 (4%)	1 (1%)	3 (4%)	-
4	CiTnI(ABC)nlacZ*	39	17 (45%)	4 (10%)	7 (20%)	11 (28%)	-
	CiTnI(ABC)nlacZ	38	26 (68%)	6 (15%)	4 (11%)	2 (5%)	30/35 (86%)
	CiTnI(AB)nlacZ*	139	70 (50%)	25 (18%)	21 (15%)	23 (16%)	23/27 (85%)
	CiTnI(AC)nlacZ	64	53 (83%)	3 (5%)	6 (9%)	2 (3%)	15/18 (83%)
	CiTnI(AC)nlacZ*	40	39 (98%)	1 (2%)	0	0	-

Each row represents a separate electroporation. Electroporations on the same batch of fertilized embryos are grouped together. Batches 1 and 2 were stained with X-gal for 2.5 hours. Batches 3 and 4 were stained with X-gal for 10 minutes.

[A] GFP transfection control is reported for electroporations evaluated for GFP expression as number of positive embryos/number of embryos examined. For Batch 1 the same co-transfected embryos were subsequently fixed and stained with X-gal. In the process some embryos were lost hence the small discrepancies between the total number of embryos evaluated for GFP expression and those stained for X-gal. For Batch 4 a subset of embryos were examined for GFP expression and returned to the main population which was then fixed and X-gal stained.

* Electroporations represented in Figure 28.

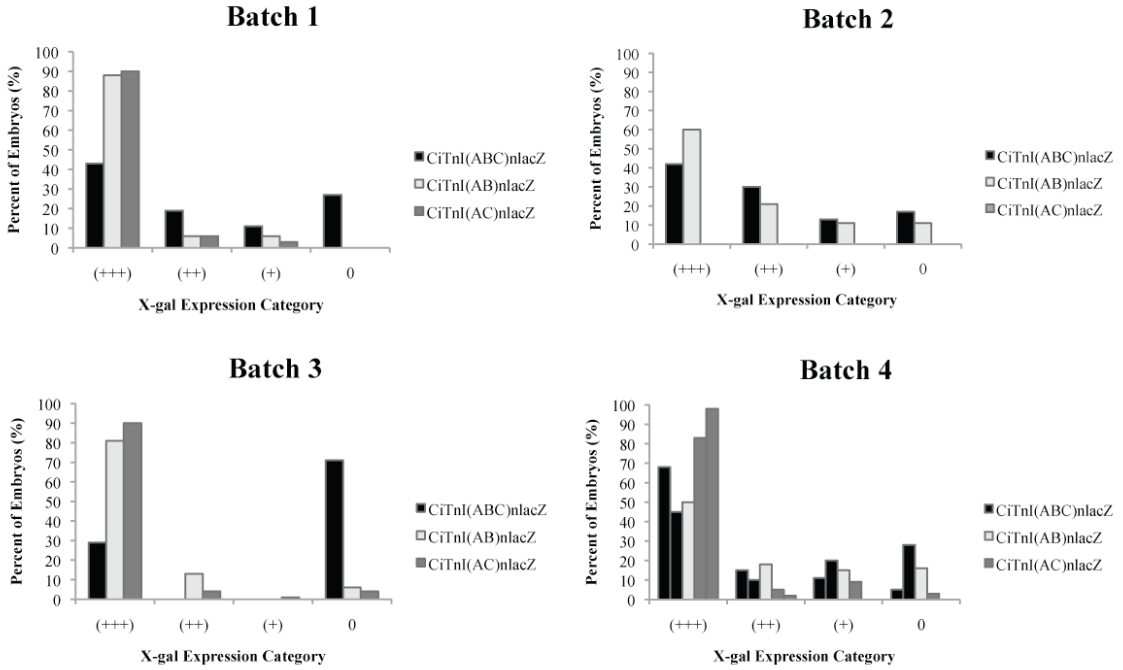


Figure 28: Histograms of the β-galactosidase expression levels of outron-retaining constructs in *Ciona* 12h tailbud embryos. Each graph represents a different batch of fertilized embryos, and each bar a separate electroporation. This is a graphical representation of the numerical data reported in Table 4.

Table 5: Expression strength ratios of outron-retaining constructs in *Ciona* 12h tailbud embryos

	CiTnI(ABC)nlacZ	CiTnI(AB)nlacZ	CiTnI(AC)nlacZ
Mortimer (2007)	0.16	13.2	-
Batch 1 of fertilized eggs	1.59	>88	>90
Batch 2 of fertilized eggs	2.47	5.45	-
Batch 3 of fertilized eggs	0.41	13.5	22.5
Batch 4 of fertilized eggs	1.61	2.71	27.7
	13.6	-	>98

Expression strength ratio is the percentage of (+++) scoring embryos divided by the percentage of (0) scoring embryos (see Table 4). If the percentage of (0) scoring embryos is zero then the ratio is reported as > percentage of (+++) embryos.

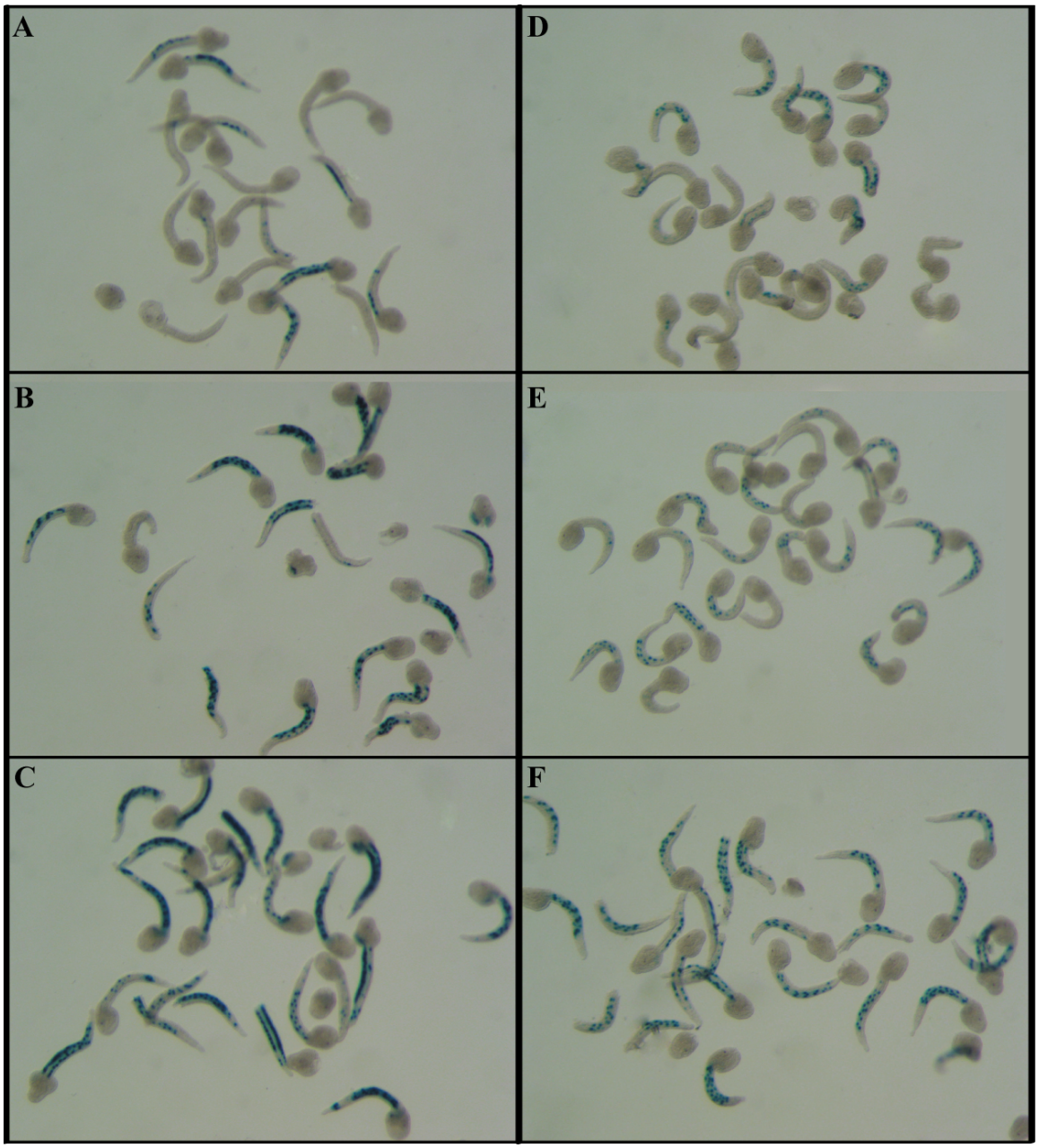


Figure 29: Expression of constructs CiTnI(ABC)nlacZ, CiTnI(AB)nlacZ and CiTnI(AC)nlacZ in *Ciona* 12h embryos.
A, B, C) 2.5 hours X-gal stain of embryos transformed with A) CiTnI(ABC)nlacZ, B) CiTnI(AB)nlacZ and C) CiTnI(AC)nlacZ, see * marked Batch 1 constructs in Table 4.
D, E, F) 10 minutes X-gal stain of embryos transformed with D) CiTnI(ABC)nlacZ, E) CiTnI(AB)nlacZ and F) CiTnI(AC)nlacZ, see * marked Batch 4 constructs on Table A.
Embryos in A, B and C appear to have stronger stained cells than E, D, F, but the number of stained cells per embryo are similar for each construct under both staining conditions.
Electroporations included 22.5 ug of the outron test construct (ABC, AB, or AC) and 2.5 ug of CiTnI(-1.5 kb)eGFP plasmid DNA.

DISCUSSION

Co-electroporation with GFP control construct

The co-electroporations with the CiTnI(-1.5 kb)eGFP construct confirmed that the differences between the β-galactosidase activity of different constructs reported by S. Mortimer was not due to different electroporation efficiencies. Therefore, my studies, by confirming that the construct with the longest retained-outron segment showed the lowest expression levels, add further support to the 5'-UTR sanitization hypothesis of the SL *trans*-splicing function. Any batch of fertilized eggs showed similar levels of GFP expression in separate electroporations revealing transfection reproducibility (see Table 4, batches 1 and 4). Therefore, any differences in β-galactosidase activity were due to differences in the DNA constructs, not the electroporation. It would be wise to include the control construct in all future electroporations to confirm nothing unusual occurs in a specific experiment and that the variability, such as that seen for the CiTnI(ABC)nlacZ construct in batch 4 is not due to a transfection irregularity.

β-galactosidase activity of outron-retaining constructs

The results of the β-galactosidase activity of the CiTnI(AC)nlacZ construct argue against the specific deleterious element hypothesis. They clearly demonstrate that there are no deleterious elements in the -177 to -79 region, section C, of the outron. The β-galactosidase expression of the CiTnI(AC)nlacZ construct was clearly much stronger than that of CiTnI(ABC)nlacZ and in all cases at least equivalent to CiTnI(AB)nlacZ. Notwithstanding any possible interactions between different parts of the outron these results strongly support the non-specific length hypothesis that the lower β-galactosidase activity seen from the CiTnI(ABC)nlacZ when compared to CiTnI(AB)nlacZ,

CiTnI(AC)nlacZ and CiTnI(A)nlacZ is due to the greater length of the retained outron per se. A plausible theory is that the translational machinery simply has a greater probability of falling off the mRNA if it must read through a long 5'-untranslated sequence before reaching the protein coding sequence.

It would follow that if longer 5'-UTRs are deleterious to expression of a reporter gene, shorter 5'-UTRs should increase the reporter activity. However, this is not observed as there does not appear to be an increase in β-galactosidase activity between CiTnI(AB)nlacZ/CiTnI(AC)nlacZ and the shorter CiTnI(A)nlacZ construct. This may be because there is a threshold 5'-UTR length below which a good level translation and therefore expression of the reporter gene will occur. It should be noted that the difference between CiTnI(ABC)nlacZ and CiTnI(AB)nlacZ/CiTnI(AC)nlacZ is only 98 bases. It would be of interest to elucidate if a threshold exists by beginning with CiTnI(AB)nlacZ and making progressively longer outron-retaining constructs by adding back region C in increments. Additionally it would be interesting for future experiments to test if another gene that exhibits SL *trans*-splicing in *Ciona* is also negatively affected by a long retained outron.

CHAPTER 5 – GENERAL DISCUSSION AND OVERALL CONCLUSIONS

Evolution of SL *trans*-splicing in the deuterostomes

The studies undertaken in the course of this thesis touched on several aspects of SL *trans*-splicing. The first main question addressed was the phylogenetic distribution of SL *trans*-splicing in the deuterostomes. As established by the lack of common sequences at the 5'-ends of different 5'-RACE cDNA molecules, none of the species carefully examined (*Branchiostoma*, *Saccoglossus* or *Strongylocentrotus*) utilize this post-transcriptional modification. Thus, my data suggests that SL *trans*-splicing is not an ancestral trait within the deuterostomes, but rather, arose *de novo* in the tunicates. This is the first *de novo* emergence of SL *trans*-splicing to be documented by rigorious 5'-RACE methodology, although the study of Douris et al. (2010), based on analysis of conventional ESTs, which would have failed to find short SLs, is also consistent with *de novo* emergence of *trans*-splicing within the tunicates. Because all tunicate groups examined, including the basal larvacean Oikopleura (Ganot et al., 2004), carry out *trans*-splicing, this mechanism must have emerged early in the tunicate lineage. Therefore, the focus of SL *trans*-splicing in the deuterostomes now shifts to how and why it evolved in the tunicates.

Possible mechanism of *de novo* SL *trans*-splicing evolution

Given the presence of *cis*-splicing, the evolution of SL *trans*-splicing only requires two additional features, an SL RNA to act as the SL donor and an unpaired acceptor site in the 5'-region of the target pre-mRNA (Nilsen, 2001). In species that *trans*-splice, the only difference between a gene that is SL *trans*-spliced and one that is not is the presence of the 5' unpaired acceptor site (Blumenthal, 1995). In a non-*trans*-

splicing species genes could readily gain and lose unpaired splice acceptor sites in the 5'-UTR without any functional consequences, as there is no suitable donor SL RNA. Thus, if an SL RNA did arise *de novo* in a non-*trans*-splicing species there would likely already be a sub-population of genes present that happened to contain unpaired splice acceptor sites, so that *trans*-splicing could occur immediately and perhaps undergo positive selection if any of these *trans*-splicing events happened to be beneficial for the organism. It has been hypothesized that the SL RNA may have evolved from the appearance of a splice-donor site on an Sm-binding U snRNA (reviewed in Nilsen, 2001). Alternatively, the SL RNA could have evolved from the promoter, first exon and part of the first intron of a protein coding gene, which already has a splice-donor site, and thus only required the appearance of an Sm-binding site (Hastings, 2005). It would be of interest to further investigate the SL RNA of *Ciona*, and perhaps in an in vitro assay introduce a splice-donor site on an Sm-binding U snRNA to examine if it will easily function in *trans*-splicing. Furthermore the consequences of the evolution SL *trans*-splicing on tunicate gene regulation also emerges as a worthy area of study, especially in comparison with other deuterostome groups that do not use this mechanism.

Role of SL *trans*-splicing in monocistronic gene expression

The second main question addressed in this thesis was if SL *trans*-splicing acted as a way to sanitize the 5'-UTR by removing deleterious elements. A prior study from our lab suggested that experimental retention of outrons of sufficient length has a deleterious effect on gene expression. My studies confirm these observations and extend them them by suggesting that it is the length of the outron, rather than the presence of specific deleterious elements, that interferes with expression. From the reporter gene assay in this study the length of the outron removed by *trans*-splicing appears to be an

important factor in the expression of the SL *trans*-spliced TnI gene. This suggests that further investigation should be focused on varying the length of retained-outrons of different *trans*-spliced genes, perhaps also in different *trans*-splicing species.

The results presented in this thesis on the phylogenetic distribution of SL *trans*-splicing in the deuterostomes and on the likely 5'-UTR sanitization function of SL trans-splicing of monocistronic genes, are useful contributions likely to influence the direction of future research.

6. REFERENCES

Ast, G. (2004). How did alternative splicing evolve? *Nat Rev Genet, 5*(10), 773-782.

Bernard, P., Gabarit, P., Bahassi, E. M., & Couturier, M. (1994). Positive-selection vectors using the F plasmid ccdB killer gene. *Gene, 148*(1), 71-74.

Blumenthal, T. (1995). Trans-splicing and polycistronic transcription in Caenohabditis elegans. *Trends in Genetics, 11*(4), 132-136.

Blumenthal, T. (2004). Operons in eukaryotes. *Briefings in Functional Genomics & Proteomics, 3*(3), 199-211.

Blumenthal, T. (2005). Trans-splicing and operons. *WormBook*, 1-9.

Blumenthal, T., & Gleason, K. S. (2003). Caenorhabditis elegans operons: form and function. *Nature Reviews Genetics, 4*, 110-118.

Boue, S., Letunic, I., & Bork, P. (2003). Alternative splicing and evolution. *BioEssays, 25*(11), 1031-1034.

Bourlat, S. J., Juliusdottir, T., Lowe, C. J., Freeman, R., Aronowicz, J., Kirschner, M., et al. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature, 444*(7115), 85-88.

Brehm, K., Jensen, K., & Frosch, M. (2000). mRNA trans-splicing in the human parasitic cestode Echinococcus multilocularis. *J Biol Chem, 275*(49), 38311-38318.

Brehm, K., Hubert, K., Sciutto, E., Garate, T., & Frosch, M. (2002). Characterization of a spliced leader gene and of trans-spliced mRNAs from Taenia solium. *Molecular and Biochemical Parasitology, 122*(1), 105-110.

Bruzik, J. P., Doren, K. V., Hirsh, D., & Steitz, J. A. (1988). Trans splicing involves a novel form of small nuclear ribonucleoprotein particles. *Nature, 335*(6190), 559-562.

Cameron, C. B., Garey, J. R., & Swalla, B. J. (2000). Evolution of the chordate body plan: New insights from phylogenetic analyses of deuterostome phyla. *Proceedings of the National Academy of Sciences, 97*(9), 4469-4474.

Campbell, D. A., Thornton, D. A., & Boothroyd, J. C. (1984). Apparent discontinuous transcription of Trypanosoma brucei variant surface antigen genes. *Nature, 311*(5984), 350-355.

Chang, Y.-F., Imam, J. S., & Wilkinson, M. F. (2007). The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annual Review of Biochemistry, 76*(1), 51-74.

Cheng, G., Cohen, L., Ndegwa, D., & Davis, R. E. (2006). The Flatworm Spliced Leader 3' -Terminal AUG as a Translation Initiator Methionine. *Journal of Biological Chemistry, 281*(2), 733-743.

Cleto, C. (2002). *Analysis of transcriptional elements of an ascidian troponin I gene.* M.Sc. thesis, McGill University, Montreal, Canada.

Cleto, C. L., Vandenberghe, A. E., MacLean, D. W., Pannunzio, P., Tortorelli, C., Meedel, T. H., et al. (2003). Ascidian Larva Reveals Ancient Origin of Vertebrate-Skeletal-Muscle Troponin I Characteristics in Chordate Locomotory Muscle. *Molecular Biology and Evolution, 20*(12), 2113-2122.

Conrad, R., Thomas, J., Spieth, J., & Blumenthal, T. (1991). Insertion of part of an intron into the 5' untranslated region of a Caenorhabditis elegans gene converts it into a trans-spliced gene. *Molecular and Cellular Biology, 11*(4), 1921-1926.

Corbo, J. C., Levine, M., & Zeller, R. W. (1997). Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, *Ciona* intestinalis. *Development, 124*(3), 589-602.

D'Alessio, J. M., & Gerard, G. F. (1988). Second-strand cDNA synthesis with E.coli DNA polymerase I and RNase H: the fate of information at the mRNA 5' terminus and the effect of E.coli DNA ligase. *Nucleic Acids Research, 16*(5), 1999-2014.

Damgaard, C. K., & Lykke-Andersen, J. (2011). Translational coregulation of 5,Ä≤TOP mRNAs by TIA-1 and TIAR. *Genes & Development, 25*(19), 2057-2068.

Davis, R. E., Hardwick, C., Tavernier, P., Hodgson, S., & Singh, H. (1995). RNA trans-splicing in flatworms. Analysis of trans-spliced mRNAs and genes in the human parasite, Schistosoma mansoni. *Journal of Biological Chemistry, 270*(37), 21813-21819.

Davis, R. E. (1996). Spliced leader RNA trans-splicing in metazoa. *Parasitology Today, 12*(1), 33-40.

Davis, R. E. (1997). Surprising diversity and distribution of spliced leader RNAs in flatworms. *Molecular and Biochemical Parasitology, 87*(1), 29-48.

Delsuc, F., Brinkmann, H., Chourrout, D., & Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. [10.1038/nature04336]. *Nature, 439*(7079), 965-968.

Derelle, R., Momose, T., Manuel, M., Da Silva, C., Wincker, P., & Houliston, E. (2010). Convergent origins and rapid evolution of spliced leader trans-splicing in metazoa: insights from the ctenophora and hydrozoa. *RNA, 16*(4), 696-707.

Doren, K. V., & Hirsh, D. (1988). Trans-spliced leader RNA exists as small nuclear ribonucleoprotein particles in Caenorhabditis elegans. [10.1038/335556a0]. *Nature, 335*(6190), 556-559.

Douris, V., Telford, M. J., & Averof, M. (2010). Evidence for multiple independent origins of trans-splicing in Metazoa. *Mol Biol Evol, 27*(3), 684-693.

Douzery, E. J. P., Snell, E. A., Bapteste, E., Delsuc, F. d. r., & Philippe, H. (2004). The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences of the United States of America, 101*(43), 15386-15391.

Drouin, G., & de Sa, M. M. (1995). The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Molecular Biology and Evolution, 12*(3), 481-493.

Eckhart, L., Ban, J., Ballaun, C., Weninger, W., & Tschachler, E. (1999). Reverse Transcription-Polymerase Chain Reaction Products of Alternatively Spliced mRNAs Form DNA Heteroduplexes and Heteroduplex Complexes. *Journal of Biological Chemistry, 274*(5), 2613-2615.

Erwin, D. H., Laflamme, M., Tweedt, S. M., Sperling, E. A., Pisani, D., & Peterson, K. J. (2011). The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science (New York, N.Y.), 334*(6059), 1091-1097.

Frohman, M. A., Dush, M. K., & Martin, G. R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences, 85*(23), 8998-9002.

Ganot, P., Kallesoe, T., Reinhardt, R., Chourrout, D., & Thompson, E. M. (2004). Spliced-Leader RNA trans Splicing in a Chordate, Oikopleura dioica, with a Compact Genome. *Molecular and Cellular Biology, 24*(17), 7795-7805.

Gasparini, F., & Shimeld, S. (2011). Analysis of a botryllid enriched-full-length cDNA library: insight into the evolution of spliced leader trans-splicing in tunicates. *Development Genes and Evolution, 220*(11), 329-336.

Gee, H. (2006). Evolution: Careful with that amphioxus. [10.1038/439923a]. *Nature, 439*(7079), 923-924.

Hamilton, T. L., Stoneley, M., Spriggs, K. A., & Bushell, M. (2006). TOPs and their regulation. *Biochemical Society transactions, 34*(Pt 1), 12-16.

Hannon, G. J., Maroney, P. A., & Nilsen, T. W. (1991). U small nuclear ribonucleoprotein requirements for nematode cis- and trans-splicing in vitro. *Journal of Biological Chemistry, 266*(34), 22792-22795.

Hannon, G., Maroney, P., Yu, Y., Hannon, G., & Nilsen, T. (1992). Interaction of U6 snRNA with a sequence required for function of the nematode SL RNA in trans-splicing. *Science, 258*(5089), 1775-1780.

Hastings, K. E. (2005). SL trans-splicing: easy come or easy go? *Trends Genet, 21*(4), 240-247

Johnson, P. J., Kooter, J. M., & Borst, P. (1987). Inactivation of transcription by UV

irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. *Cell, 51*(2), 273-281.

Khare, P., Mortimer, S. I., Cleto, C. L., Okamura, K., Suzuki, Y., Kusakabe, T., et al. (2011). Cross-validated methods for promoter/transcription start site mapping in SL trans-spliced genes, established using the *Ciona* intestinalis troponin I gene. *Nucleic Acids Research, 39*(7), 2638-2648.

Kiss, T., Fayet, E., Jady, B. E., Richard, P., & Weber, M. (2006). Biogenesis and Intranuclear Trafficking of Human Box C/D and H/ACA RNPs. *Cold Spring Harbor Symposia on Quantitative Biology, 71*, 407-417.

Knoll, A. H. (1992). The Early Evolution of Eukaryotes: A Geological Perspective. *Science, 256*(5057), 622-627.

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene, 361*(0), 13-37.

Krause, M., & Hirsh, D. (1987). A trans-spliced leader sequence on actin mRNA in C. elegans. *Cell, 49*(6), 753-761.

Lall, S., Friedman, C. C., Jankowska-Anyszka, M., Stepinski, J., Darzynkiewicz, E., & Davis, R. E. (2004). Contribution of Trans-splicing, 5'-Leader Length, Cap-Poly(A) Synergism, and Initiation Factors to Nematode Translation in an Ascaris suum Embryo Cell-free System. *Journal of Biological Chemistry, 279*(44), 45573-45585.

Lasda, E. L., Allen, M. A., & Blumenthal, T. (2010). Polycistronic pre-mRNA processing in vitro: snRNP and pre-mRNA role reversal in trans-splicing. *Genes Dev, 24*(15), 1645-1658.

Lee, M. G.-S., & Van der Ploeg, L. H. T. (1997). Transcription of Protein-Coding genes in Trypanosomes by RNA Polymerase I. *Annual Review of Microbiology, 51*(1), 463-489.

Lemaire, P. (2011). Evolutionary crossroads in developmental biology: the tunicates. *Development, 138*(11), 2143-2152.

Liou, R. F., & Blumenthal, T. (1990). trans-spliced Caenorhabditis elegans mRNAs retain trimethylguanosine caps. *Molecular and Cellular Biology, 10*(4), 1764-1768.

Loh, E. Y., Elliott, J. F., Cwirla, S., Lanier, L. L., & Davis, M. M. (1989). Polymerase chain reaction with single-sided specificity: analysis of T cell receptor delta chain. *Science (New York, N.Y.), 243*(4888), 217-220.

Lynch, M., & Kewalramani, A. (2003). Messenger RNA Surveillance and the Evolutionary Proliferation of Introns. *Molecular Biology and Evolution, 20*(4), 563-571.

MacLean, D. W., Meedel, T. H., & Hastings, K. E. M. (1997). Tissue-specific Alternative Splicing of Ascidian Troponin I Isoforms: Redesign of a Protein Isoform-generating Mechanism During Chordate Evolution. *Journal of Biological Chemistry, 272*(51),

32115-32120.

Mair, G., Ullu, E., & Tschudi, C. (2000). Cotranscriptional Cap 4 Formation on the Trypanosoma brucei Spliced Leader RNA. *Journal of Biological Chemistry, 275*(37), 28994-28999.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics, 9*, 387-402.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380.

Maroney, P. A., Denker, J. A., Darzynkiewicz, E., Laneve, R., & Nilsen, T. W. (1995). Most mRNAs in the nematode Ascaris lumbricoides are trans-spliced: a role for spliced leader addition in translational efficiency. *RNA, 1*(7), 714-723.

Maruyama, K., & Sugano, S. (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene, 138*(1,Äì2), 171-174.

Matera, A. G., Terns, R. M., & Terns, M. P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol, 8*(3), 209-220.

Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L., et al. (1999). Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Research, 27*(6), 1558-1560.

Mattick, J. S. (2004). RNA regulation: a new genetics? *Nat Rev Genet, 5*(4), 316-323.

Matsumoto, J., Dewar, K., Wasserscheid, J., Wiley, G. B., Macmil, S. L., Roe, B. A., et al. (2010). High-throughput sequence analysis of *Ciona* intestinalis SL trans-spliced mRNAs: Alternative expression modes and gene function correlates. *Genome Research, 20*(5), 636-645.

Mortimer, S. (2007). *Experimental analysis of trans-splicing of an ascidian troponin I gene.* M.Sc. thesis, McGill University, Montreal, Canada.

Nilsen, T. W. (1993). Trans-Splicing of Nematode Premessenger RNA. *Annual Review of Microbiology, 47*(1), 413-440.

Nilsen, T. W. (1994). RNA-RNA interactions in the spliceosome: unraveling the ties that bind. *Cell, 78*(1), 1-4.

Nilsen, T. W. (2001). Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends in Genetics, 17*(12), 678-680.

Northcutt, R. G. (2012). Evolution of centralized nervous systems: Two schools of evolutionary thought. *Proceedings of the National Academy of Sciences, 109*(Supplement 1), 10626-10633.

Orom, U. A., Nielsen, F. C., & Lund, A. H. (2008). MicroRNA-10a Binds the 5' UTR of Ribosomal Protein mRNAs and Enhances Their Translation. *Molecular Cell, 30*(4), 460-471.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., & Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annual Review of Biochemistry, 55*(1), 1119-1150.

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M., et al. (2007). A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research, 35*(19), e130.

Parsons, M., Nelson, R. G., Watkins, K. P., & Agabian, N. (1984). Trypanosome mRNAs share a common 5' spliced leader sequence. *Cell, 38*(1), 309-316.

Patel, S. B., & Bellini, M. (2008). The assembly of a spliceosomal small nuclear ribonucleoprotein particle. *Nucleic Acids Research, 36*(20), 6482-6493.

Pellizzoni, L., Yong, J., & Dreyfuss, G. (2002). Essential Role for the SMN Complex in the Specificity of snRNP Assembly. *Science, 298*(5599), 1775-1779.

Perry, K. L., Watkins, K. P., & Agabian, N. (1987). Trypanosome mRNAs have unusual "cap 4" structures acquired by addition of a spliced leader. *Proceedings of the National Academy of Sciences, 84*(23), 8190-8194.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., et al. (2011). Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature, 470*(7333), 255-258.

Plank, T.-D. M., & Kieft, J. S. (2012). The structures of nonprotein-coding RNAs that drive internal ribosome entry site function. *Wiley Interdisciplinary Reviews: RNA, 3*(2), 195-212.

Pouchkina-Stantcheva, N. N., & Tunnacliffe, A. (2005). Spliced Leader RNA-Mediated trans-Splicing in Phylum Rotifera. *Molecular Biology and Evolution, 22*(6), 1482-1489.

Poole, A. M., Jeffares, D. C., & Penny, D. (1998). The Path from the RNA World. *Journal of Molecular Evolution, 46*(1), 1-17.

Rajkovic, A., Davis, R. E., Simonsen, J. N., & Rottman, F. M. (1990). A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. *Proceedings of the National Academy of Sciences, 87*(22), 8879-8883.

Romfo, C. M., Maroney, P. A., Wu, S., & Nilsen, T. W. (2001). 3' splice site recognition in nematode trans-splicing involves enhancer-dependent recruitment of U2 snRNP. *RNA, 7*(6), 785-792.

Ronaghi, M., Uhlen, M., & Nyren, P. l. (1998). A Sequencing Method Based on Real-Time

Pyrophosphate. *Science, 281*(5375), 363-365.

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual (Second Edition) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Satoh, N. (2003). The ascidian tadpole larva: comparative molecular development and genomics. *Nat Rev Genet, 4*(4), 285-295.

Satoh, N., & Jeffery, W. R. (1995). Chasing tails in ascidians: developmental insights into the origin and evolution of chordates. [doi: 10.1016/S0168-9525(00)89106-4]. *Trends in Genetics, 11*(9), 354-359.

Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A., & Satoh, N. (2005). An Integrated Database of the Ascidian, *Ciona* intestinalis: Towards Functional Genomics. *Zoological Science, 22*(8), 837-843.

Satou, Y., Hamaguchi, M., Takeuchi, K., Hastings, K. E. M., & Satoh, N. (2006). Genomic overview of mRNA 5'-leader trans-splicing in the ascidian *Ciona* intestinalis. *Nucleic Acids Research, 34*(11), 3378-3388.

Satou, Y., Mineta, K., Ogasawara, M., Sasakura, Y., Shoguchi, E., Ueno, K., et al. (2008). Improved genome assembly and evidence-based global gene model set for the chordate *Ciona* intestinalis : new insight into intron and operon populations. *Genome Biology, 9*(10), R152.

Seto, A. G., Zaug, A. J., Sobel, S. G., Wolin, S. L., & Cech, T. R. (1999). Saccharomyces cerevisiae telomerase is an Sm small nuclear ribonucleoprotein particle. *Nature, 401*(6749), 177-180.

Shuman, S. (2002). What messenger RNA capping tells us about eukaryotic evolution. *Nat Rev Mol Cell Biol, 3*(8), 619-625.

Sierro, N., Li, S., Suzuki, Y., Yamashita, R., & Nakai, K. (2009). Spatial and temporal preferences for trans-splicing in Ciona intestinalis revealed by EST-based gene expression analysis. *Gene, 430*(1-2), 44-49.

Sonenberg, N., & Hinnebusch, A. G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell, 136*(4), 731-745.

Stover, N. A., & Steele, R. E. (2001). Trans-spliced leader addition to mRNAs in a cnidarian. *Proc Natl Acad Sci U S A, 98*(10), 5693-5698.

Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., & Sugano, S. (1997). Construction and characterization of a full length-enriched and a 5'end-enriched cDNA library. *Gene, 200*(1,2), 149-156.

Suzuki, Y., Yamashita, R., Nakai, K., & Sugano, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Research, 30*(1), 328-

331.

Thomas, J. D., Conrad, R. C., & Blumenthal, T. (1988). The C. elegans trans-spliced leader RNA is bound to Sm and has a trimethylguanosine cap. *Cell, 54*(4), 533-539.

Thompson, J. R., Marcelino, L. A., & Polz, M. F. (2002). Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by reconditioning PCR. *Nucleic Acids Research, 30*(9), 2083-2088.

Turbeville, J. M., Schulz, J. R., & Raff, R. A. (1994). Deuterostome phylogeny and the sister group of the chordates: evidence from molecules and morphology. *Molecular Biology and Evolution, 11*(4), 648-655.

Vandenberghe, A. E., Meedel, T. H., & Hastings, K. E. M. (2001). mRNA 5'-leader trans-splicing in the chordates. *Genes & Development, 15*(3), 294-303.

Will, C. L., & Luhrmann, R. (2001). Spliceosomal UsnRNP biogenesis, structure and function. *Current opinion in cell biology, 13*(3), 290-301.

Yeats, B. (2009). *Spliced leader (SL) trans-splicing in the ascidian tunicate* Ciona intestinalis*: molecular characterization of the SL RNA*. M.Sc. thesis, McGill University, Montreal, Canada.

Yeats, B., Matsumoto, J., Mortimer, S. I., Shoguchi, E., Satoh, N., & Hastings, K. E. M. (2010). SL RNA Genes of the Ascidian Tunicates *Ciona* intestinalis and *Ciona* savignyi. *Zoological Science, 27*(2), 171-180.

Zayas, R. M., Bold, T. D., & Newmark, P. A. (2005). Spliced-Leader trans-Splicing in Freshwater Planarians. *Molecular Biology and Evolution, 22*(10), 2048-2054.

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., et al. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences, 104*(11), 4618-4623.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., & Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques, 30*(4), 892-897.

Zorio, D. A., & Blumenthal, T. (1999). U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in Caenorhabditis elegans. *RNA, 5*(4), 487-494.

Zorio, D. A. R., Cheng, N. N., Blumenthal, T., & Spieth, J. (1994). Operons as a common form of chromosomal organization in C. elegans. *Nature, 372*(6503), 270-272.

APPENDIX I – OLIGONUCLEOTIDES

Oligonucleotide Number	**Sequence (5' – 3')**
9200	GAATTCTACCTCAGAGGAGTCATATNNNNNN
9038	GAATTCTACCTCAGAGGAGTCATAT
GeneRacer 5' Primer (Invitrogen)	CGACTGGAGCACGAGGACACTGA
GeneRacer 5' Nested Primer (Invitrogen)	GGACACTGACATGGACTGAAGGAGTA
9015	GGATCCGATTCTATTTGAATAAG
9394	GTAATACGACTCACTATAGGGCGAATTGGTACC
9395	AGTTCCTCTTCAGATTTGTCCTAGAAGAATAACTT
9396	AAGTTATTCTTCTAGGACAAATCTGAAGAGGAACT
9397	AAATCATCATTAAAGCGAGTGGCAACATGGAAATC
9411 (MID-2)	CCATCTCATCCCTGCGTGTCTCCGACTCAG***ACGCTCGACA***ggacactgacatggactgaaggagta
9412 (MID-3)	CCATCTCATCCCTGCGTGTCTCCGACTCAG***AGACGCACTC***ggacactgacatggactgaaggagta
9413 (MID-4)	CCATCTCATCCCTGCGTGTCTCCGACTCAG***AGCACTGTAG***ggacactgacatggactgaaggagta
9414 (MID-11)	CCATCTCATCCCTGCGTGTCTCCGACTCAG***TGATACGTCT***ggacactgacatggactgaaggagta
9415 (MID-7)	CCATCTCATCCCTGCGTGTCTCCGACTCAG***CGTGTCTCTA***ggacactgacatggactgaaggagta
9416 (MID-10)	CCATCTCATCCCTGCGTGTCTCCGACTCAG***TCTCTATGCG***ggacactgacatggactgaaggagta
9418 (MID-1)	CCATCTCATCCCTGCGTGTCTCCGACTCAG***ACGAGTGCGT***ggacactgacatggactgaaggagta
9417	CCTATCCCCTGTGTGCCTTGGCAGTCTCAGgaattctacctcagaggagtcatat

For oligos 9411 – 9418: upper case bases correspond to the 454 GS-FLX Titanium sequence (including 4-base key), bold and italics upper case bases correspond to the unique MID tag and lower case bases correspond to the template specific sequence (the GeneRacer 5' Nested Primer sequence for oligos 9411 – 9416, 9418 and the 3'-anchor-based reverse primer sequence for oligo 9417).

APPENDIX II – CONSTRUCT NOMENCLATURE

Construct	**Previous Nomenclature**	**Origin**
CiTnI(-1.5 kb)nlacZ	CiTnInuclacZ(-1.5)	Cleto, 2002
pB(CiTnI –1.5)	-	Levasseur, this thesis
pEGFP-N1	-	Clontech
CiTnI(-1.5 kb)eGFP	-	Levasseur, this thesis
CiTnI(ABC)nlacZ	CiTnI(-838/-79)nlacZ	Mortimer, 2007
CiTnI(AB)nlacZ	CiTnI(-838/-177)nlacZ	Mortimer, 2007
CiTnI(AC)nlacZ	-	Levasseur, this thesis
CiTnI(A)nlacZ	CiTnI(-838/-262)nlacZ	Mortimer, 2007

APPENDIX III – INITIAL SANGER SEQUENCING OF *CIONA* SL AND OLIGO-CAPPED cDNA PRODUCTS

SL PCR CLONES

List of 18 clones sequence ID and their corresponding matches to the KH *Ciona* gene model set genome (Satou et al., 2005). If sequence did not match a KH gene model, the top alignment to the NCBI EST database [*Ciona* (taxid: 7718)] is listed.

Sequence ID	***Ciona* KH gene model name/ NCBI EST accession number**
>ID|8206781 SL3_P1104273_031	KH.C2.744.v1.A.SL1-1
>ID|8206749 SL4_P1104273_032	KH.C3.88.v2.A.SL3-1
>ID|8206760 SL6_P1104273_048	KH.C9.640.v1.A.SL1-1
>ID|8206761 SL7_P1104273_063	65 base, no matches
>ID|8206771 SL8_P1104273_064	KH.L132.7.v1.C.SL1-1
>ID|8206772 SL9_P1104273_079	DC993977.1
>ID|8206784 SL10_P1104273_080	KH.C5.661.v1.A.SL1-1
>ID|8206785 SL11_P1104273_095	FK056686.1
>ID|8206791 SL12_P1104273_096	KH.C6.200.v2.A.SL2-1
>ID|8206792 SL13_P1104273_013	KH.C12.99.v1.A.SL3-1
>ID|8206736 SL14_P1104273_014	KH.C3.968.v2.A.SL5-1
>ID|8206737 SL15_P1104273_029	KH.C5.661.v1.A.SL1-1
>ID|8206747 SL16_P1104273_030	KH.S739.1.v1.A.SL1-1
>ID|8206748 SL17_P1104273_045	KH.C8.539.v1.A.SL1-1
>ID|8206758 SL18_P1104273_046	KH.C9.216.v1.A.SL1-1
>ID|8206759 SL19_P1104273_061	KH.L57.9.v1.B.SL1-1
>ID|8206769 SL20_P1104273_062	KH.C9.216.v1.A.SL7-1
>ID|8206770 SL21_P1104273_077	KH.C11.204.v1.A.SL2-1

OLIGO-CAPPED PCR CLONES

Both forward and reverse orientations of the insert seen due to non-directional TA cloning. pBluescript SK II + vector sequence has been removed.

GeneRacer 5' Nested Primer sequence – bold and underlined
Random Reverse Primer Anchor Sequence – bold
5'-end of RNA oligonucleotide cap sequence – highlighted
SL sequence – underlined

14 reads representing bona fide cDNA molecules:

>ID|8200658 Oligo96_GB121-HASTNGSTN1_082.ab1
ggacctgactggactgaaggagtaaaaatctgttttctgcagtaccaaaaatgaacaagtcagcactcctt
ctccttctgctcatcggacttcttgtcctcaccgaatcttcggttgggttgtttggtcgtcgtcgtcgtcg
ctggtggcatcgcagacgttatgaccaacagacaatggaacaagatgaggacatggagttccaagaattgg

aagaagacgatcatccaagtaacaaataataagtgaacaagttgaagtcggatgattttaataaattgatc
ag

EST match (97% Identity): BW649092.1, Hypothetical protein

>ID|8200652 Oligo84_GB121-HASTNGSTN1_084.ab1
gattctcctcgaggagtcatatggccgctgatcaatttatta
aaatcatccgacttcaacttgttcacttattgtgtgttacttggatgatcgtcttcttcaatttcttggaa
ctccatgtcctcatcttgttccattgtctgttggtcataacgtctgcgatgccaccagcgacgacgacgac
gaccaaacaacccaaccgaagattcggtgaggacaagaagtccgatgagcagaaagagaaggagtgctgac
ttgttcatttttggtactgcagaaaacagattttctactccttcagtccatgtcagtgtcctcgtgctcct
ttctactccttcagtccatgtcagtgtccccgtgctcctttctactccttcagtccatgtcagtgtcctcg
tgctcctttctactccttcagtccatgtcagtgtcctcgttttctactccttcagtccatgtcagtgtcct
cgttttctactccttcagtccatgtcagtgtcctcgtattc***tactccttcagtccatgtcagtgtcc***

EST match (97% Identity): BW649092.1, Hypothetical protein

>ID|8200649 Oligo78_GB121-HASTNGSTN1_036.ab1
ggacctgactggactgaaggagtagaaaatctgttttctgcagtaccaaaaatgaacaagtcagcactcct
tctccttctgctcatcggacttcttgtcctcaccgaatcttcggttgggttgtttggtcgtcgtcgtcgtc
gctggtggcatcgcagacgttatgaccaacagacaatggaacaagatgaggacacggggttccaagaattg
gaagaagacgatcatccaagtaacaaataataagtgaacaagttgaagtcggatgattttaataaattgat
cagcggc***atatgactcctctgaggtagaattc***

EST match (97% Identity): BW649092.1, Hypothetical protein

NOTE: Above 3 reads represent the same mRNA.

>ID|8200639 Oligo58_GB121-HASTNGSTN1_072.ab1
ggacactgactggactgaaggagtagaaaccagtcttctacaggacacaaagtaaacatgaacaagtcagc
tcttctccttttgctcgtcgtcggactcctcgttctcactgacaccgccgttgccggtcgtcgtcgcagat
ggtggaaccaaaagccagtaaatgctgaagacaccgatgcgatgatggactctgtttggggcc***atatgact
cctctgaggtagaattc***

EST match (96% Identity): BW649454.1, Hypethetical protein

>ID|8200622 Oligo24_GB121-HASTNGSTN1_094.ab1
ggacctgactggactgaaggagtagaaaccagtcttctacaggacacaaagtaaacatgaacaagtcagct
cttctccttttgctcgtcgtcggactcctcgttctcactgacaccgccgttgccggtcgtcgtcgcagatg
gtggaaccaaaagccagtaaatgctgaagacaccgatgcgatgatggactctgtttggggcc***atatgactc
ctctgaggtagaattc***

EST match (96% Identity): BW649454.1 Hypothetical protein

NOTE: Above 2 reads represent the same mRNA.

>ID|8200582 Oligo91_GB121-HASTNGSTN1_049.ab1
ggtcgctaacttcatctctgactgcatccatcaactcatcgttgtcttcagcattaactggcttttggttg
taccatctgcgacgacgacctgcaacggcggtgtcagtgagaacgaggagtccgacgacgagcaaaaggag
aagagctgacttgttcatgtttactttgtgtcctgtagaagactggttttc***tactccttcagtccatgtca
gtgtcc***

EST match (99% Identity): AJ227672.1, Hypothetical protein

>ID|8200654 Oligo88_GB121-HASTNGSTN1_018.ab1
ggacctgacatggactgaaggagtagaaatacactttaaaagctttaacgagtatccattggagggcaagt
ctggtgccagcagccgcggtaattccagctccaaaagtatatatttaagttgttgcggttgaaaagctcgt
agttggattttgggcgcgggcggtcggtccgtcgcgaggcgtgtactggtcgacccggccttacgtccggt

tctccggcggtgctcttgactgagtgtcggcggcggccggaaagtttactttgaaaaaattagagtgttca
aagcaggctctcgagcctggataatggtgcatggaataatggaataggacctcggttctattttgttggtt
ttcggagcgcgaggtaatgattaagagggacagacgggggg***atatgactcctctgaggtagaattc***

EST match (99% Identity): FF825635.1, cleft lip and palate
transmembrane protein 1 homolog

>ID|8200651 Oligo82_GB121-HASTNGSTN1_068.ab1
ggacactgaatggactgaaggagtagaaacttttactacacgccgtgcgaaaatgggtaaacccagtggt
cttcgttgtgcccgaaagttgaagaacagacgtcgtgatcagaaatggcacgataaggattacaaaaaatc
tcacttgggaacagctttgaaagccaatccttttggaggagcttcacatgcaaagggaatcgtacttgaaa
agattggtgttgaagctaaacaaccaaactcagctattagaaaatgtgtccgagtg***atatgactcctctgg
gtagaattc***

EST match (98% Identity): FF907257.1, 40S ribosomal protein S23

>ID|8200564 Oligo75_GB121-HASTNGSTN1_019.ab1
gaggagtagaaattctatttgaataaaggcgtatagttgggtaagtataagttttataattgtgctgcgta
taaaccacacaatttggggaggccaaaacagaatgcaagttactgctgctgctgccaatattcctatctgc
cactttaaacggtgataacaagggaagggaaatatttacaacatcgaca***atatgactcctctgaggtagaa
ttc***

SL present. No *Ciona* EST matches. Aligns with KH Genome (94% Identity):
KhL44 from +119960 to +120120

>ID|8200647 Oligo74_GB121-HASTNGSTN1_004.ab1
tctccccgtcccgttatcccccgccgcgcgtccgcccgcacgcccannngacgcggttagattcgcctggt
atttactcgaggagtcatattggccccaaacagagtccatcatcgcatcggtgtcttcagcatttactggc
ttttggttccaccatctgcgacgacgaccggcaacggcggtgtcagtgagaacgaggagtccgacgacgag
caaaaggagaagagctgacttgttcatgtttactttgtgtcctgtagaagactggttttc***tactccttcag
tccatgtcagtgtcc***

EST match (96% Identity): AJ227672.1, Hypothetical protein

>ID|8200642 Oligo64_GB121-HASTNGSTN1_022.ab1
atggacgaaggagtagaaacttttcgaaattgttttgaccaaaggctcaccgcngnttcttgaggtcaaga
ttacgtcggcctatcacaaaatacaatcaacatgagtgcttacaaagacgcnnngaaaggtcccacagggc
gagcaggaagtacagatccataaaatccggattacgctgacttcccgtaacgtccagagcttggagaagg***a
tatgactcctctgaggtagaattc***

EST match (97% Identity): BW650079.1, 40S ribosomal protein S20

>ID|8200617 Oligo49_GB121-HASTNGSTN1_007.ab1
ggacctgactggactgaaggagtagaaagtaattccagctccaaagtatatatttaagttgttgcggttga
aaagctcgtagttggattttgggcgcgggcggtcggtccgtcgcgaggcgtgtactggtcgacccggcctt
acgtccggttctccggcggtgctcttgactgagtgtcggcggcggccggaaagtttactttgaaaaaatta
gagtgttcaaagcaggctctcgagcctggataatggtgcatggaataatggaataggacctcggttctatt
ttgttggttttcggagcgcgaggtaatgattaagagggacagacgggggc***atatgactcctctgaggtaga
attc***

EST match (99% Identity): FF895708.1, Hypothetical protein

>ID|8200596 Oligo45_GB121-HASTNGSTN1_073.ab1
ggacctgactggactgaaggagtagaaaccagtcttctacaggacacaaagtaaacatgaacaagtcagct
cttctccttttgctcgtcgtcggactcctcgttctcactgacaccgccgttgccggtcgtcgtcgcagatg
gtacaaccaaaagccagttaatgctgaagataacgatgagttgatggatgcagtcagagatgaagttatgc
aacatctcagccaagcgg***atatgactcctctgaggtagaattc***

EST match (99% Identity): AJ227672.1, Hypothetical protein

>ID|8200630 Oligo40_GB121-HASTNGSTN1_026.ab1
aggagagaaatcgttcatgcgtatcgtcgtgtgtgagagcatagataccttacaaggtatcgatngtgaga
ggggcgtgttgtaggcgtttttacagttaaaaatggtacgaaactaagagccaggctgttgtgtgggga***at atcactcctctgaggtagaattc***

EST match (97% Identity): FK184818.1

58 Primer-multimer products:
Examples of typical primer-multimer products

>ID|8200595 Oligo57_GB121-HASTNGSTN1_071.ab1
ggacactgaatggactgaaggagtagaaagactggagcacga***ggacactgacatggactgaaggagta***gaa
agactggagcacga***ggacactgacatggactgaaggagta***ga***atatgactcctctgaggtagaattc***

>ID|8200641 Oligo62_GB121-HASTNGSTN1_006.ab1
aaggatagaaa***ggacactgacatggactgaaggagta***gaaa***ggacactgacatggactgaaggagta***gaaa
ggacactgacatggactgaaggagtagaaa***ggacactgacatggactgaaggagta***gaag***ggacactgaca tggactgaaggagta***gaaacga***ggacactgacatggactgaaggagta***gaaacga***ggacactgacatggac tgaaggagta***gaaacga***ggacactgacatggactgaaggagta***gaaacga***ggacactgacatggactgaag gagta***gaaacga***ggacactgacatggactgaaggagta***gaaacga***ggacactgacatggactgaaggagta***
gaaacga***ggacactgac/atatgactcctctgaggtagaattc***

APPENDIX IV – SANGER SEQUENCING OF TA CLONED OLIGO-CAP PCR PRODUCTS

A) Clones from libraries made by subjecting early-cycle PCR products to gel filtration column chromatography to remove any small products and then utilized to do a full PCR with this DNA as the template. All PCRs done with GeneRacer 5' Nested Primer and reverse primer (9038)

Row	Clone name	Insert length	Accession number of NCBI EST alignment
1	*Petromyzon* adult DNA 3	671	FD720824.1
2	*Petromyzon* adult DNA 4 *	378	No significant alignment
3	*Petromyzon* adult DNA 5	889	EE743067.1
4	*Petromyzon* adult DNA 6 *	> 221	EG334238.1
5	*Petromyzon* adult DNA 7	484	FD708473.1
6	*Petromyzon* adult DNA 8	680	DY253687.1
7	*Petromyzon* adult DNA 9	857	FD720879.1
8	*Petromyzon* adult DNA 10	> 803	CO548571.1
9	*Ciona* DNA 1	374	FF698470.1
10	*Ciona* DNA 2	498	BW649337.1
11	*Ciona* DNA 3	373	FK158342.1
12	*Ciona* DNA 4	418	FF949602.1
13	*Ciona* DNA 5	607	FF882591.1
14	*Ciona* DNA 6	365	FF766340.1
15	*Ciona* DNA 7	758	FF721568.1
16	*Ciona* DNA 8	385	BW649986.1
17	*Petromyzon* embryo DNA 1 *	95	No significant alignment
18	*Petromyzon* embryo DNA 2	No insert	N/A
19	*Petromyzon* embryo DNA 3	246	DY250291.1
20	*Petromyzon* embryo DNA 4	111	EE741494.1
21	*Petromyzon* embryo DNA 5 *	237	No significant alignment
22	*Petromyzon* embryo DNA 6	220	FD708384.1
23	*Petromyzon* embryo DNA 7	275	DY252904.1
24	*Petromyzon* embryo DNA 8	61	No significant alignment
25	*Branchiostoma* embryo DNA 1	721	FE593273.1
26	*Branchiostoma* embryo DNA 2	367	FE556651.1
27	*Branchiostoma* embryo DNA 3	486	BW711497.1
28	*Branchiostoma* embryo DNA 4	570	BW718596.1
29	*Branchiostoma* embryo DNA 5	> 852	FE561531.1
30	*Branchiostoma* embryo DNA 6	> 872	BW888761.1
31	*Branchiostoma* adult DNA 1	541	FE580945.1
32	*Branchiostoma* adult DNA 2	462	CF919377.1
33	*Branchiostoma* adult DNA 3	> 297	BW953159.1
34	*Branchiostoma* adult DNA 4	668	No significant alignment
35	*Branchiostoma* adult DNA 5	396	BI381368.1
36	*Branchiostoma* adult DNA 6	441	BW877907.1
37	*Branchiostoma* adult DNA 7	506	FE564313.1
38	*Branchiostoma* adult DNA 8	> 148	FE594264.1
39	*Saccoglossus* embryo DNA 1	175	Primer-multimer
40	*Saccoglossus* embryo DNA 2	459	FE646030.1
41	*Saccoglossus* embryo DNA 3	379	FF122688.1

42	*Saccoglossus* embryo DNA 4	> 951	CO550231.1
43	*Strongylocentrotus* embryo DNA 1	> 941	EC434930.1
44	*Strongylocentrotus* embryo DNA 2	605	CX686267.1
45	*Strongylocentrotus* embryo DNA 3	526	EC430971.1
46	*Strongylocentrotus* embryo DNA 4	315	CX685023.1
47	*Strongylocentrotus* embryo DNA 5	470	CX696597.1
48	*Strongylocentrotus* embryo DNA 6	699	EC429082.1

Insert lengths of > # indicates sequencing reaction ended before reverse primer.
* *Petromyzon* sequences showing simple repeats similar to those found in 454 reads (see Figure 25A).

B) *Ciona* clones made by 5 cycles, gel chromatography, then 30 cycles with the original primers, then an additional 15 cycles with the 454 primer.

GeneRacer 5' Nested Primer sequence – bold and underlined
Random Reverse Primer Anchor Sequence – bold
5'-end of RNA oligonucleotide cap sequence – highlighted

DNA 9

NCTATCNNCTGTGTGCCTTGGCAGTCTCAGGAATTCTACCTCAGAGGAGTCATATC TACTCCTTCAGTCCA TGTCAGTGTCCTCGTGCTCCTTTC***TACTCCTTCAGTCCATGTCAGTGTCC***TCGTGCTCCTTTC***TACTCCTT CAGTCCATGTCAGTGTCC***TCGTGCTCCTTTC***TACTCCTTCAGTCCATGTCAGTGTCC***TCGTGCTCCTTTC***T ACTCCTTCAGTCCATGTCAGTGTCCACGCACTCGTCTGAGTCGGAGACACGCA***

DNA 10

NNTATCCCCTGTGTGCCTTGGCAGTCTCAGGAATTCTACCTCAGAGGAGTCATATGTCCCCGTGCTCCTTT C***TACTCCTTCAGTCCATGTCAGTGTCC***TCGTGCTCCTTTC***TACTCCTTCAGTCCATGTCAGTGTCC***TCGTG CTCC***TTTC***TACTCCTTCAGTCCATGTCAGTGTCCACGCACTCGTCTGAGTCGGAGACA***

DNA 11

NCTATCCCCTGTGTGCCTTGGCAGTCTCAGGAATTCTACCTCAGAGGAGTCATATGCTCCAGTTTC***TACTC CTTCAGTCCATGTCAGTGTCC***TCGTGCTCCAG***TTTC***TACTCCTTCAGTCCATGTCAGTGTCC***TCGTGCTCC AG***TTTC***TACTCCTTCAGTCCATGTCAGTGTCCACGCACTCGTCTGAGTCGGAGACACGCAGGGATGAGATG G***

DNA 12

NNTATCCCCTGTGTGCCTTGGCAGTCTCAGGAATTCTACCTCAGAGGAGTCATATTCTCGTG***TTTC***TACTC CTTCAGTCCATGTCAGTGTCC***TCGTG***TTTC***TACTCCTTCAGTCCATGTCAGTGTCC***TCGTG***TTTC***TACTCC TTCAGTCCATGTCAGTGTCCACGCACTCGTCTGAGTCGGAGACACGCAGGGATGAGATGG***

DNA 13

CCTATCCCCTGTGTGCCTTGGCAGTCTCAGGAATTCTACCTCAGAGGAGTCATATGTGCTCCAGTC***TTTC***T ACTCCTTCAGTCCATGTCAGTGTCC***TCGTGCTCCAGTC***TTTC***TACTCCTTCAGTCCATGTCAGTGTCC***TCG TGCTCCAGTC***TTTC***TACTCCTTCAGTCCATGTCAGTGTCCACGCACTCGTCTGAGTCGGAGACACG***

C) Clones from libraries made by 30 cycle PCR with 454 adapter primers immediately after early-cycle gel filtration

Row	Clone name	Insert length	Accession number of NCBI EST alignment
1	*Ciona* DNA 14	277	BW314884.1
2	*Ciona* DNA 15	538	BW314884.1
3	*Petromyzon* adult DNA 11	669	DY250031.1
4	*Petromyzon* adult DNA 12	517	EC384197.1
5	*Petromyzon* adult DNA 13 *	191	No significant alignment
6	*Petromyzon* embryo DNA 9	411	EG025104.1
7	*Petromyzon* embryo DNA 10	392	CO549189.1
8	*Branchiostoma* embryo DNA 7	677	FE571704.1
9	*Branchiostoma* embryo DNA 8	491	FE544005.1
10	*Branchiostoma* adult DNA 9	744	FE579846.1
11	*Branchiostoma* adult DNA 10	557	FE594882.1
12	*Saccoglossus* embryo DNA 5	640	FF672700.1
13	*Saccoglossus* embryo DNA 6	753	FF483477.1
14	*Strongylocentrotus* embryo DNA 7	542	CX556816.1
15	*Strongylocentrotus* embryo DNA 8	428	CX680056.1

APPENDIX V – VERIFICATION OF 454 SEQUENCING OF CONTROL *CIONA* cDNA POPULATION

GeneRacer 5' Nested Primer sequence – bold and underlined
Random Reverse Primer Anchor Sequence – bold
5'-end of RNA oligonucleotide cap sequence – highlighted
SL sequence – underlined

Bona fide 5' cDNAs:

30 reads, mean read length: 217 bases

>HNSZ03401CY9OP length=197 xy=1104_2871 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACTTTCTTCAGCGACCGAACTGGTTTGCTGTTCATAATTGTG
GGATAAATGGCTTGCGCACGGCCATTAGTGTCCGTGTATTCGGAAAGGGTGTTGCTTCGGGTCCATTACAT
TGCCAGCAGTTTTAAGGCACCGATCCGTCCT***ATATGACTCCTCTGAGGTAGAATT***

>HNSZ03401BM8GQ length=227 xy=0557_2312 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAAGTAAACAT
GAACAAGTCAGCGCTTCTCCTTTGCTCGTCGTCGGACTCCTCGTTCTCACTGACACCGCC
GTTGCCGGTCGTCGTCGCAGATGGTGGAATCAAAGCCAGTAAATGCTGAAGACACCGATG
CGATGATGGACTCTGTTGGGGCC***ATATGACTCCTCTGAGGTAGAATT***

>HNSZ03401A1D71 length=252 xy=0308_3247 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAAGTAAACAT
GAACAAGTCAGCTCTTCTCCTTTGCTCGTCGTCGGACTCCTCGTTCTCACTGACACCGCC
GTTGCCGGTCGTCGTCGCAGATGGTACAACCAAAGCCAGTTAATGCTGAAGACAACGATG
AGTTGATGGATGCAGTCAGAGATGAAGTTATGCAACATCTCAGCCAAGCGGA***ATATGACT***
CCTCTGAGGTAG

>HNSZ03401BU6RT length=211 xy=0648_0631 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAATCTGTTTCTGCAGTACCAAAATGAACAAGT
CAGCACTCCTTCTCTTTCTGCTCATCGGACTTCTTGTCCTCACCGAATCTTCGGTTGGTT
GTTTGGTCGTCGTCGTCGTCGCTGGTGGCATCGCAGACGTTATGACCAACAGACAATGGA
ACAAGATGAGGACATGGAGTTCCAAGAATTG

>HNSZ03401EQXNH length=145 xy=1829_3659 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAATTCTATTTGAATAAGGTATTGCCTTCTACG
TAGGCGAATAGGTTAGTTGTTAAACGAGAACCTCTGGCAAACTGCTACTCTAAGCCACGC
G***ATATGACTCCTCTGAGGTAGAATT***

>HNSZ03401CEC1X length=206 xy=0866_2307 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACATTACTACCAGGACGAACACTAACGAACA
AGATGAATTCTACTACTTTACTTCTGCGCACTTTGTCTTATACTGTCAAGTCCAGCATTG
TTACCGCTGGTGCTGCAAGTCCTTCCTGAGTTACGGAATGTTCGATGTTAGGTAGAGGAA
CAGTGCGTTCGAGTTCAGCTGTGTAC

>HNSZ03401COQV5 length=376 xy=0984_3475 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACTTTCGAAAAACTTTATGATATCTGGACGC

ATCGAGGACTTCTAGTGTCAATATATATATAACTGAATATGTCTGACGGTGAAGGAGATG
ATGTTCCGGTCGCAGTACAACCCAGCGGTCCTATGGACCTTAATACAGCGCTTCAGGAGG
TTCTAAAACAGCAGTCATTAACGACGGTCTTGCCCGCGGTCTGAACGATGCGCAAGGCCT
TGGATAACGCCAGGCTCATCTTTGCGTTCTTGCTAACACTGTGACGACCAGCTATGTTAA
CTCATCGAGGCTTTGTGCATGAGCATCAATCAGTTTGATCAAGTTGACGACAGCAATAGC
TCGGGATGGGCTGGTC

>HNSZ03401ARTT6 length=231 xy=0199_3388 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAAGTAAACAT
GAACAAGTCAGCTCTTCTCCTTTGCTCGTCGTCGGACTCCTCGTTCTCACTGACACCGCC
GTTGCCGGTCGTCGTCGCAGATGGTACAACCAAAGCCAGTTAATGCTGAAGACAACGATG
AGTTGATGGATGCAGTCAGAGATGAAGTTATGCAACATCTCAGCCAAGCGG

>HNSZ03401BZMOO length=198 xy=0698_3078 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACTTTTCAGTGCGTCGGCCCAAGACCGAAGC
TACCTGCACATTGTCCATGTGAGCGAGCTCAAGGAGCGATAAATATGGGATTGTAAAGTT
GTAAAACGAAGCGTACTTCAAGCGGTACCAAGTAAATATCGAAGACGACGAGAAGGCGTG
ACCGATTATTTGTCGAGG

>HNSZ03401EB07O length=226 xy=1660_0658 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAAGTAAACAT
GAACAAGTCAGCTCTTCTCCTTTGTCTCGTCGTCGGACTCCTCGTTCTCACTGACACCGC
CGTTGCCGGTCGTCGTCGCAGATGGTGGAACCAAAGCCAGTAAATGCTGAAGAACACCGA
TGCGATGATGGACTCTGTTGGGGC***ATATGACTCCTCTGAGGTAGAA***

>HNSZ03401CYUEL length=196 xy=1099_3547 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACTTTTCAGTGCGTCGGCCCCAAGACCGAAG
CTACCTGCACATTGTCCATGTGAGCGAGCTCAAGGAGCGATAAATATGGGGATTTGTAAA
GTTGTAAAAACGAAGCGTACTTCAAGCGGTACCAAGTAAATATCGAAGACGACGAGAAGG
CGTGACCGATTATTTG

>HNSZ03401BUGNH length=240 xy=0639_3643 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACATTACTACCAGGACGAAACACAACGAACA
AGATGAATTCTCTACTTTACTTCTGCGCACTTTGTCTTATACTGTCAAGTCCAGCATTGG
TTACCGCTGGTGCTGCAAGTCCTTTCCTGAGTTACGGAAATGTTCGATGTTAGGTAGAGG
AACAGTGCATTCGAGTTCAGCTGTGTACTTGCTAATGGAACTGCTGACTGCATGGACAAA

>HNSZ03401CZXSD length=299 xy=1112_1339 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGGTCGAAACTTTCCATTTCGTCGCCAAAATGGTTGCCCG
AAATAAGAAACCAGAAAGCTTCGTGGGCACGTAAGCCACGGTCATGGACGTGTAGGCAAA
CACCGCAAGCATCCTGGAGGTCGAGGTAATGCTGGTGGACAGCATCATCATAGAATCAAC
TATGATAATACCATCCTGGTTACTTGGTAAGTTGGTATGAGACATTTCCATTTGAAGAGA
ACCCACAGCTTGCCCAATCATCAACCTTGATCGATGTGGACGCTTGTCAGTGAGCAACC

>HNSZ03401AN15B length=195 xy=0156_3661 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACTTTTCAGTGCGTCGGCCCAAGACCGAAGC
TACCTGCACATTGTCCATGTGAGCGAGCTCAAGGAGCGATAAATATGGGGATTTGTAAAA
GTTGTAAAAACGAAGCGTACTTCAAGCGGTACCAAGTAAATATCGAAGACGACGAGAAGG
CGTGACCGATTATTT

>HNSZ03401AJMCO length=176 xy=0106_1366 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTGGAAAATTCTATTAGAATAAGAAATAATTTAAAGT
GGAATTTTCCGTAAGTTATAAAGAGATATCGAGCGATGGATCGTCAAGGCAAAGCTAACA
ACGACAACAAGTCAAACCAATGCAACCCAAAC***ATATGACTCCTCTGAGGTAGAATT***

>HNSZ03401CEXW6 length=175 xy=0873_0680 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTGGAAATTCTATTAGAATAAGAAATAATTTAAAGTG
GAATTTTCCGTAAGTTATAAAGAGATATCGAGCGATGGATCGTCAAGGCAAAGCTAACAA
CGACAACAAGTCAAACCAATGCAACCCAAAC***ATATGACTCCTCTGAGGTAGAATT***

>HNSZ03401CKA1I length=229 xy=0934_1108 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACTTCTGATAAAACATGTCACGTTACGGAGG
AAACGATTTAGAGGCTCTAAGGGTGAAGCTCAGTTCGACACCTTCAGTCAACAAAAGATA
GAAAACTTGACCGACCAAACAAGAATATAATGGACCACCTGGATGGACGCCGATGGCTTG
ATTGAGACCAACTATGACAAATAGTTGAATCGTTGATGACATGTGCCTC

>HNSZ03401D4FY7 length=196 xy=1573_2897 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAAGTAAACGT
GAACAAGTCAGCTCTTCTCCTTTTGCTCGTCGTCGGACTCCTCGTTCTCACTGACACCGC
CGTTGCCGGTCGTCGTCGCAGATGGTGGAACCAAAACCAGTAAATGCTGAAGACACCGAT
GCGATGATGGACTCTG

>HNSZ03401EH6U4 length=222 xy=1730_1198 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAAGTAAACAT
GAACAAGTCAGCTCTTCTCCTTTTGCTCGTCGTCGGACTCCTCGTTCTCACTGACACCGC
CGTTGCCGGTCGTCGTCGCAGATGGTGGAACCAAAGCCAGTAAATGCTAAGACACCGATG
CGATGATGGACTCTGTTTGGGC***ATATGACTCCTCTGAGGTAG***

>HNSZ03401A4V2E length=234 xy=0348_2500 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTACGAAACTTTTACGTGTGTAGTGTTGTGCAGCCCA
CACGGCACAATGAAGATCATACACTCAAGTCAAGTTGTGACTTTGCCCGACGAAGTTAAA
GTTACTGTTAAAGGACGAGTGGTTACAGTTCGTGGCCAAGAGGTGTACTTAAAAAAATTC
AACCATTGAAGGTTGAGCTGACAAAGTTGGTACCAACTCTAGTAAAGTGGACAA

>HNSZ03401BV5JQ length=90 xy=0659_0644 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAACTCTTACCAGCTGGGAGAAGTGCATGTGCGT
TGCGTCGAAATTAGATGGATTTCTAAGTTT

>HNSZ03401CD96N length=105 xy=0865_2685 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAACCAGTCTTCTACAGGACACAAGTAACATGAA
CAAGTCAGCTCTTCTCCTTTGTCTCGTCGTCGGACGTCCTCGTTC

>HNSZ03401DH3MK length=204 xy=1319_0842 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACATTACTACCAGGACGAACACTAACGAACA
AGATGAATTCTACTACTTACTTCTGCGCACTTGTCTTATACTGTCAAGTCCAGCATTGTT
ACCGCTGGTGCTGCAAGTCCTTCCTGAGTTACGGAATGTTCGATGTTAGGTAGAGGAACA
GTGCGTTCGAGTTCAGCTGTGTAC

>HNSZ03401DXTPH length=144 xy=1498_1299 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAATTCTATTGAATAAGGTATTGCCTTCTACGT
AGGCGAATAGGTTAGTTGTTAAACGAGAACCTCTGGCAAACTGCTACTCTAAGCCACGCG
ATATGACTCCTCTGAGGTAGAATT

>HNSZ03401EAK5J length=289 xy=1643_2821 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAATCTGTTTCTGCAGTACCAAAATGAACAAGT
CAGCACTCCTTCTCCTTCTGCTCATCGGACTTCTTGTCCTCACCGAATCTTCGGTTGGTT
GTTTGGTCGTCGTCGTCGTCGCTGGTGGCATCGCAGACGTTATGACCAACAGACAATGGA
ACAAGATGAGGACACGGAGTTCCAAGAATTGGAAGAAGACGATCATCCAAGTACAATATA
AGTGACAGTTGAAGTCGGATGATTTATAATTGATCAGCGGT***ATATGACT***

>HNSZ03401DL9NV length=206 xy=1366_2777 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACATTACTACCAGGACGAACACAACGAACAA
GATGAATTCTACTACTTTACTTCTGCGCACTTTGTCTTATACTGTCAAGTCCAGCATTGG
TTACCGCTGGTGCTGCAAGTCCTTCCTGAGTTACGGAATGTTCGATGTTAGGTAGAGGAA
CAGTGCGTTCGAGTTCAGCTGTGTAC

>HNSZ03401EH1LC length=205 xy=1728_2558 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACATTACTACCAGGACGAACACTAACGAACA
AGATGAATTCTACTACTTACTTCTGCGCACTTGTCTTATACTGTCAAGTCCAGCATTGTT
ACCGCTGGTGCTGCAAGTCCTTCCTGAGTTACGGAATGTTACGATGTTAGGTAGAGGAAC
AGTGCGTTCGAGTTCAGCTGTGTAC

>HNSZ03401DBZWX length=376 xy=1249_2815 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAACTTTGGATTGAAGTCTTGAATGGCTCGTATC
AAAGCCAAGGAACTGCGTGGAGAGTCGAAGGACGAGCTGCTTAAGCAGCTAAACGACTTC
AAGTAGGAATTATCTACCCTGCGGGTGGCGAAGTTACCGGTGGAGCCGCGTCAAAGCTCT
CAAATATGCTTGGTCCGCAAAGCATCGCTCGTGTCCTCACGGTCATTAACCAGACTCAAA
GGATACTTGAGGAATTGTTAGGAAGAACATAAGCCAGGATTACGGCCAAAAGACCAGAGC
CTGCGTCGTCGTTTAACAAATGAGGAAGTTGAATCCGTAAAGCTTGAAAAGCAGACTATA
CCACAGCGACAGTTGC

>HNSZ03401DI1HP length=229 xy=1329_3771 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAGTAACATGA
ACAAGTCAGCTCTTCTCCTTTGCTCGTCGTCGGACTCCTCGTTCTCACTGACACCGCCGT
TGCCGGTCGTCGTCGCAGATGGTACAACCAAAGCCAGTTAATGCTGAAGACAACGATGAG
TTGATGGATGCAGTCAGAGATGAAGTTATGCAACATCTCAGCCAAGCGG

>HNSZ03401CBZY3 length=243 xy=0839_2637 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAACCAGTCTTCTACAGGACACAAGTAACATGA
ACAAGTCAGCTCTTCTCCTTTGCTCGTCGTCGGACTCCTCGTTCTCACTGACACGACTGT
TGGTTATTATCGTCGTCGTCGCAGATGGCGATGGAACCAAAGCCAGTAATGCTGAAGACA
ACGATGAACTGATGGGTGAATTAAGAGATGAAGTTATGCAACAATCAGCAAGCGG***ATATG
ACT***

Primer-multimers example:

8 reads, mean read length of 115 bases

>HNSZ03401BW6EZ length=122 xy=0670_3369 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAAA***GAGCACGA******GGACACTGACATGGACTGAAGG***
AGTAGAAA***GAGCACGA******GGACACTTACATGGACTGAAGG/ATATGACTCCTCTGAGGTAGAA***
TT

Primer-dimers examples:
38 reads, mean read length of 50 bases

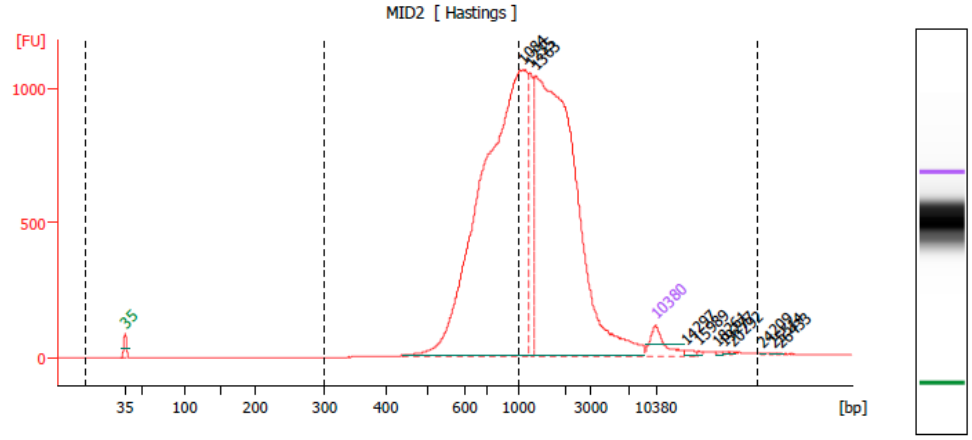
>HNSZ03401D99PE length=52 xy=1640_0272 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGAA/TATGACTCCTCTGAGGTAGAATT

>HNSZ03401D1N1J length=47 xy=1542_0357 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAAGGAGTAGA/TATGACTCCTCTGAGGTAG

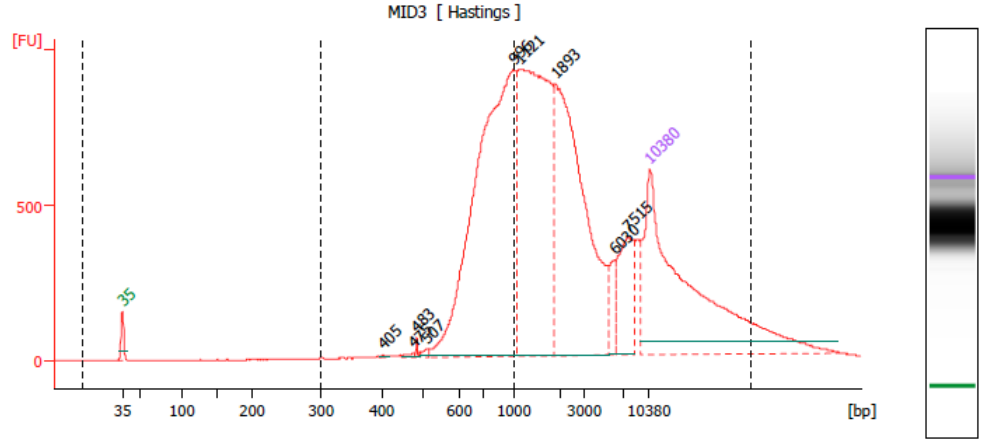
>HNSZ03401DY1YS length=43 xy=1512_1314 region=1
run=R_2012_05_09_16_56_53_
GGACACTGACATGGACTGAGGAGTA/TATGACTCCTTGAGGTAG

APPENDIX VI - BIOANALYZER ANALYSIS OF OLIGO-CAP 5'-RACE AMPLICON LIBRARIES BY MCGILL UNIVERSITY AND GENOME QUEBEC INNOVATION CENTRE

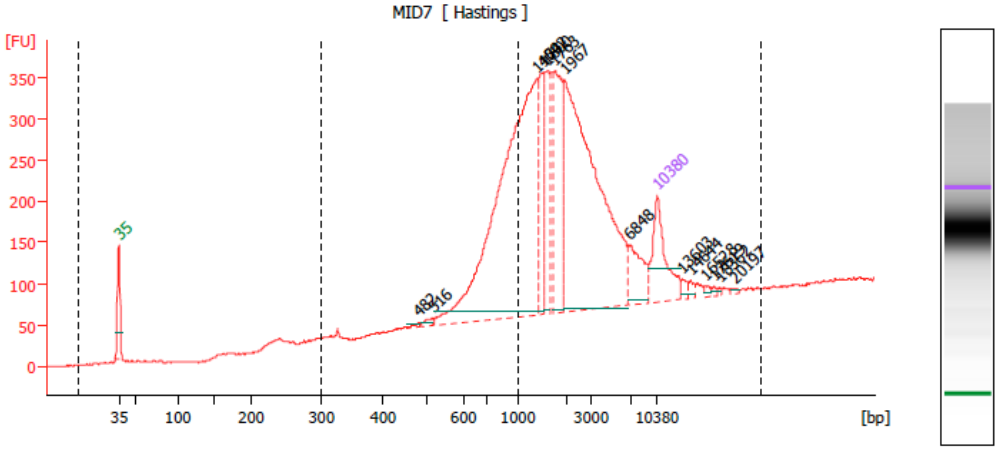
Branchiostoma Adult



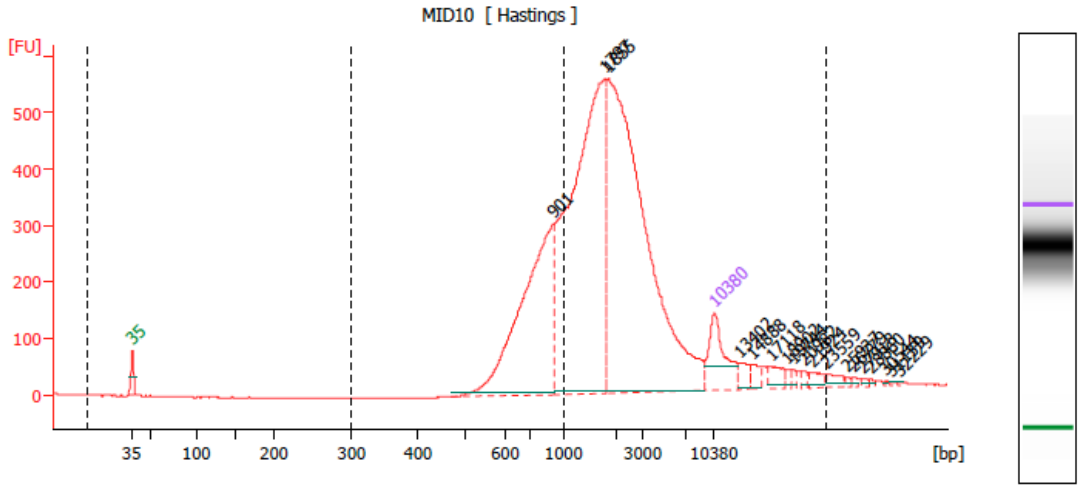
Branchiostoma Embryo



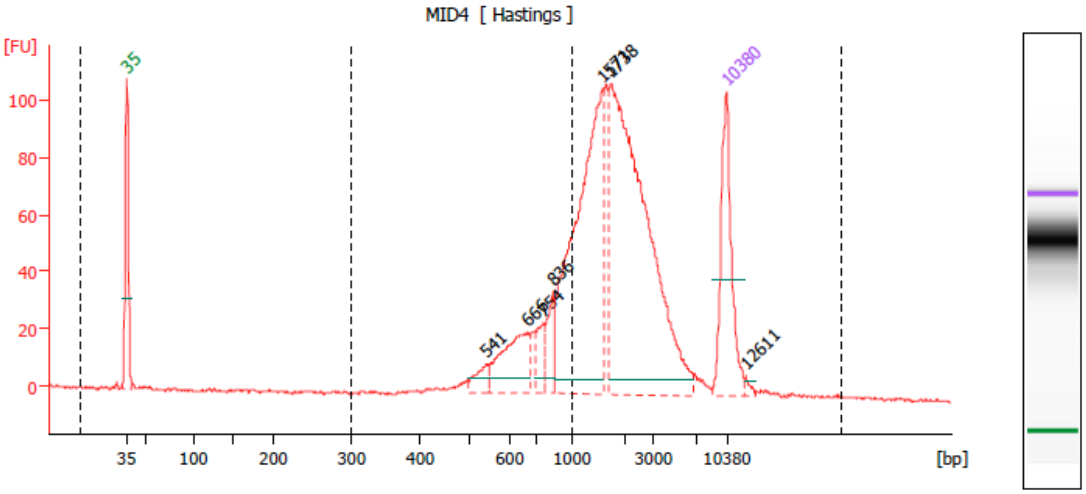
Saccoglossus Embryo



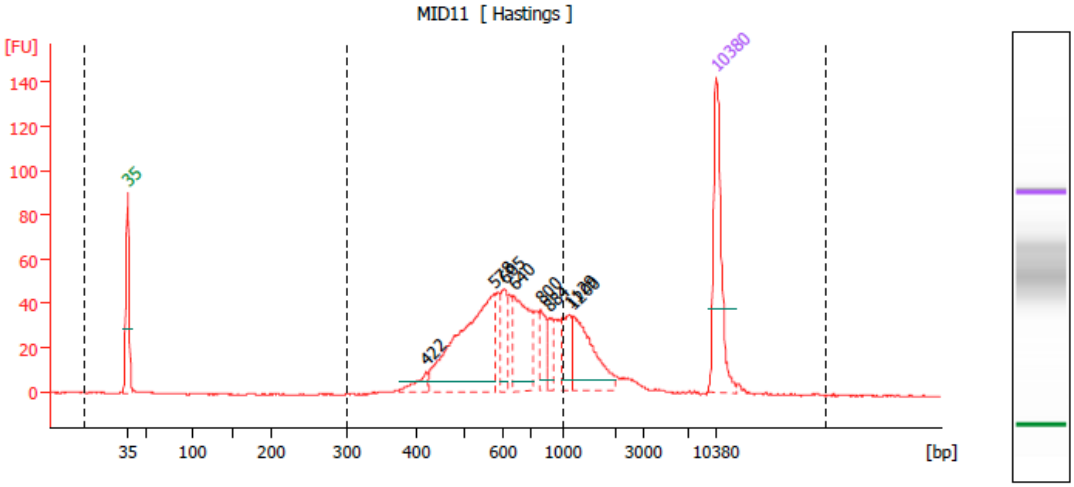
Strongylocentrotus Embryo



Petromyzon Adult



Petromyzon Embryo



APPENDIX VII – MANUAL ANALYSIS OF *BRANCHIOSTOMA* ADULT MID-2 LIBRARY

Table sorted by sequence text in alphabetical order. Only showing first 30 bases of the trimmed sequence, length of complete trimmed sequence is recorded in fourth column. Highlighted rows are ribosomal proteins and translational factors (as assessed by their gene name). Most begin with the 5' TOP sequence, as expected for 5'-complete 5'-RACE products.

Unigene gene name	**Accession No. of top hit EST**	**No. of reads**	**Trimmed length bases**	**5'-sequence of sole, or representative, read**
programmed cell death protein 5	XM_002588690.1	2	464	CAAAAAATTCACGAAGATGGCAGACCCAGA
hypothetical protein	XM_002593362.1	1	537	CAAAAAGGTTTGAAAATGAGTCGGTCGTTC
hypothetical protein	XM_002585514.1	2	518	CAAACAAAAGCCAGCTATGGTTAGATGCTA
methyltransferase-like protein 21A	BW777654.1	1	414	CACACCTTGGGGCGCCAGTTCTGGCCAGTT
hypothetical protein	XM_002597310.1	1	250	CACATTCCTGACTATCGTACGAAGAGTCAA
prefoldin subunit 4	BI381570.1	3	474	CACTGAAAAAAATCCAAACATGGCTACCGT
hypothetical protein	XM_002594824.1	2	513	CACTGACATGGACTGAAGGAGTAGAAAAAG
chymotrypsin-like protease	XM_002610330.1	4	727	CACTTTGCTAAAAAACTTCGTCACAACAGG
hypothetical protein	XM_002587275.1	1	954	CAGAAAGGTTCGCTACTGGTTTGGAAGAAG
hypothetical protein	XM_002590266.1	3	385	CAGTTCGTTTGTGACTTCTGTAGGTATACG
hypothetical protein	XM_002603773.1	1	477	CATACTCCAGAAGACCGAAGCAAGATGGGG
chymotrypsinogen B precursor	XM_002603259.1	1	93	CATCACTCGGATCCCCGATTCTACAACGTG
chymotrypsinogen B precursor	XM_002603258.1	1	531	CATCACTCGGATCCCCGATTCTACAACGTG
eukaryotic translation initiation factor 4H isoform 2	BW803355.1	6	432	CATTCTGAGTTCGCTCTCGCCGTGGAAGGA
hypothetical protein	BW841405.1	2	254	CATTTTCCTGCTACACCGCAGCGCATTTTG
putative RNA-binding protein 3	FE594155.1	1	400	CATTTTTCTACAGCGACTTGTGACTCGATT
MORN repeat-containing protein 5	XM_002609310.1	1	452	CCAAAAACAAAAAATGGAGTACACCGGTAG
hypothetical	XM_002594870.1	1	445	CCAAAACGGTCTCCAAGGAAAGCCAGAAGC

protein				
hypothetical protein	XM_002599169.1	1	590	CCAACTACACTGCTCCTGAGAACGTCCAAG
ATP synthase subunit beta, mitochondrial precursor	XM_002595071.1	2	560	CCAGTCAGCCATGTTGCGTACGGCAGCTCG
high mobility group protein B1	FE585194.1	3	521	CCATAGGCTTCGCTGCAGAACTTTGAGGTG
actin, alpha skeletal muscle	BW698205.1	1	87	CCATCTACATCTAACGGCTTTTGGTGCTTG
hypothetical protein	BW813635.1	2	702	CCCAAGATGGCGGCAGGTGAGAGAAAAAGA
actin, alpha cardiac muscle 1	XM_002610755.1	1	592	CCCATAGCTCCAGGCATCTAACTGAAGGTC
hypothetical protein	XM_002600623.1	24	491	CCCATCTACTCGTACATCCAACGGTGAGAG
hypothetical protein	BW842130.1	1	405	CCCCTTTGACCTTCGTCTCGAGAGATAGAT
troponin I, fast skeletal muscle	XM_002586492.1	2	475	CCCTAAATTTTAAACAGTCCGTCCTCTGGA
hypothetical protein	BI379595.1	2	504	CCCTTCATCGCGTCAACCGGCAGCCGCGCC
hypothetical protein	XM_002608746.1	3	550	CCGACCGCCATCTTGCTCCGTCCCGGTGAC
tropomyosin alpha-3 chain isoform 1	BW815236.1	7	546	CGAGTGGCAAGACTTGCGCACATTCCTCTG
hypothetical protein	XM_002601413.1	1	612	CGTAAATTCTAAAATTCACTCCCAGTTCAC
H1 histone family, member O, oocyte-specific-like	BW749499.1	1	804	CGTTGAAAAGTTCATTTTTGCTGTCCTTGG
T-complex protein 1 subunit alpha isoform a	XM_002609959.1	2	743	CTAGCCAAAGATGTCGGCCCTAACTGTGAT
cathepsin L2 preproprotein	XM_002611706.1	2	719	CTCACTCGATTCAACTCCTCACTCAGACAA
nascent polypeptide-associated complex subunit alpha isoform b	FE566540.1	1	674	CTCCAAACTCGCGTCATCTTTTCGTGAAAA
hypothetical protein	BW697474.1	1	458	CTCCATACACTGACTGACCCCTCCACTATC
dynein light chain 2, cytoplasmic	XM_002611762.1	1	667	CTCCCAACAAAGACACTCACAGTGACAACC
dickkopf related protein-3	AY608670.1	2	730	CTCCCACCTCTACCCATTAAGAATCGTTCC

hypothetical protein	BW699083.1	4	655	CTCCGTTTTTTGCCAGGCATCTTAGCAGGA
transcribed locus	BW709409.1	3	560	CTCTCACCTCCACCCATCGTTAATCGTTCC
60S ribosomal protein L22 proprotein	XM_002614005.1	1	674	CTCTCTCCCACTGCCCATCATGGCGCCCGC
40S ribosomal protein S4, X isoform	XM_002599030.1	1	832	CTCTCTCCCTCGGCATCCAACATGGCGCGA
*60S ribosomal protein L32	XM_002592101.1	4	729	CTCTTCCGCCATCTTGGATCCTGGAAAAAG
40S ribosomal protein S15	XM_002594975.1	1	550	CTCTTCGACTCGTACCTGCTCAAAAATGGC
60S ribosomal protein L31 isoform 1	XM_002611659.1	3	570	CTCTTCGCCATTTTCCAAGATGGCTCCTAC
elongation factor 1-alpha 1	XM_002604678.1	1	543	CTCTTCTACTCGTAGCGCGGCGCTGTGAGC
60S acidic ribosomal protein P1	BW695553.1	1	516	CTCTTTACGCCATCTTGCGCAAAAGTGTGG
**60S ribosomal protein L28 isoform 2	XM_002593652.1	15	711	CTCTTTCATTTTCCGACGAGACAGAAAAGC
60S ribosomal protein L23a	XM_002605598.1	2	519	CTCTTTCCAAAAATCATGCCGCCTAAGAAG
60S ribosomal protein L4	XM_002592359.1	1	816	CTCTTTCCCGTGTGGCGGCCATTGCTGCAC
40S ribosomal protein S24 isoform a	XM_002595985.1	1	420	CTCTTTCCCTGTGGCCCGAAAATCAGGCAA
40S ribosomal protein S24 isoform a	XM_002595987.1	1	420	CTCTTTCCCTGTGGCCCGAAAATCAGGCAA
40S ribosomal protein S20 isoform 2	XM_002593553.1	3	573	CTCTTTCCGCCAAACTTGCAAGTGGAACCA
60S ribosomal protein L19	XM_002589037.1	6	1078	CTCTTTCCGGGCTGGCAAAAAAGGAGGCAG
elongation factor 1-delta isoform 2	XM_002603074.1	4	682	CTCTTTCGCAAAGTGCCAGCCATCTTGGAA
40S ribosomal protein S21	BI377058.1	4	384	CTCTTTCGCTGTCCGCCATCTTTGGGAAGC
60S ribosomal protein L17 isoform a	XM_002601586.1	1	617	CTCTTTCGTCGGTGGGTTCAGGTAACTGAC
40S ribosomal protein S13	BW862535.1	4	601	CTCTTTCTACCAGCCGCCATCATGGGTCGT
40S ribosomal protein S14	BW775014.1	2	37	CTCTTTTCCTGGACTTCGCCATCTTGTGAG
60S ribosomal protein L15	XM_002598543.1	1	419	CTCTTTTCCTTCTGTCCGCCATCTTGGCGA
40S ribosomal protein S15a	XM_002607166.1	5	680	CTCTTTTCGCCGACATCTTGGATCGGTAAC

40S ribosomal protein S11	XM_002608356.1	2	511	CTCTTTTTCCTCTCATCGAAAAAATGGCGG
nucleoside diphosphate kinase B isoform a	XM_002598239.1	1	551	CTCTTTTTCTCCCCGGCCGCCATCTTGTGC
60S ribosomal protein L37	XM_002606002.1	2	564	CTCTTTTTTCCTTCCCACGTCTGAAAAGAT
actin, gamma-enteric smooth muscle isoform 1 precursor	FE591717.1	1	905	CTGCAGTCCAAGACTTGCCAGGTTCTTCTC
28S ribosomal protein S14, mitochondrial	XM_002604677.1	2	506	CTGGAAATAAAAAAAATGGCCGCCTCAGTTT
hypothetical protein	XM_002589806.1	2	411	CTGTTTGAACTACCCTGGGAACGTGACGGC
hypoxanthine-guanine phosphoribosyltransferase	XM_002588877.1	1	362	CTGTTTTGAAACTACCCCTGGGAACGTGAC
ferritin heavy chain	XM_002599072.1	1	630	CTTCGCCAAGTTCTTCAGCCACCAGTCGGA
ferritin heavy chain	XM_002590964.1	2	654	CTTCTCATTAGATCACAGCTCGGCTACTGA
cathepsin L2 preproprotein	FE589103.1	2	106	CTTCTGACTTGGCTTGGGAAGACGTGCAAC
hypothetical protein	BW739537.1	2	796	CTTCTTACTGACATCTCATCATGAGCCAGT
hypothetical protein	XM_002610889.1	3	707	CTTGACTCTCTGATAATCTGAGCTTTACAG
hypothetical protein	XM_002595970.1	2	677	CTTTCACCTTTCCTGTAGTAGATTAAGACT
chymotrypsin B1	XM_002602353.1	2	676	CTTTCTTCCATCGAGCGTTGTCTTACACGT
40S ribosomal protein S5	XM_002609272.1	2	868	CTTTTCCCTCCTCCATCTTGGACAGTAAGC
40S ribosomal protein S27	XM_002591230.1	2	735	CTTTTCGGCCGACGTGAGACAGTCAGCCAA
60S ribosomal protein L13 isoform 1	XM_002604882.1	1	616	CTTTTCGTCTGCGGACATCTTGCCGTAGAT
40S ribosomal protein S16	XM_002608433.1	5	703	CTTTTCTCCATAGTCGCCGTACGAAAAGGA
60S ribosomal protein L36	XM_002602234.1	1	125	CTTTTTCGGTCTCCATCTTGGATCCAGGCA
elongation factor 1-alpha 1	XM_002604676.1	1	818	CTTTTTTCAACCGAGATCTGCTTCCGTTCA
60S ribosomal protein L10a	XM_002607793.1	1	95	CTTTTTTCCTGCCTTCGTCTGCTAACGAGA
60S ribosomal protein L34	XM_002596102.1	7	459	CTTTTTTCGCCATTTTGCACCAAAAGGTTT
40S ribosomal protein S6	FE568591.1	2	660	CTTTTTTCGGCCACCGTCTCATGCGAGGCG
60S ribosomal protein L14	XM_002604832.1	1	587	CTTTTTTTCCGCCATCTTGCTGAAAAAATG

hypothetical protein	XM_002596557.1	1	813	CTTTTTTTTCTCTCCAACCTACGCTGGACG
40S ribosomal protein S10	XM_002608198.1	2	550	CTTTTTTTTTCCCGTGGCTGCCATCTTGGC
NFU1 iron-sulfur cluster scaffold homolog, mitochondrial isoform 1	XM_002613703.1	1	904	GAAAGAAAAAAACAAAGATGGCTGCCCGTG
transcribed locus	BW700784.1	5	516	GAACAAGAAGAACATCACTCAAGGCAGACA
cell division control protein 6 homolog	XM_002598373.1	2	547	GAACACAAGCTAAAAATTTTGGCGGCACGG
charged multivesicular body protein 2a	XM_002595768.1	1	707	GAAGCGCCTGTCACCAAAAGGCCACCAGTA
40S ribosomal protein S8	XM_002602289.1	1	470	GAATCCCTTGCAACTTCCTCTTTCCAGCTC
hypothetical protein	FE574579.1	1	790	GACAAACTTCTTCCCCATGGTTCTTCAGTT
transcribed locus	BW693503.1	1	623	GACAGTCAAAACCTGACAGCTGATAGGTGC
fucolectin-1-like	BW706283.1	1	643	GAGAAATTTAGACCCGGGAGCTGTGTACTG
hypothetical protein	BW699829.1	1	523	GATACTAGTACGCGACAGAAAAAAGCCGTA
NADH dehydrogenase [ubiquinone] iron-sulfur protein 7, mitochondrial precursor	BW705390.1	2	752	GATCCAACATGGCGGCGTTGGGATTCCTGC
allograft inflammatory factor 1 isoform 3	BW880949.1	2	396	GATTCGAGAAAAACTTTGCCCGCGGGGTTA
thioredoxin domain-containing protein 12 precursor	XM_002602520.1	1	95	GATTTCATCTGCTCTGAGGAACCCAAATGG
hypothetical protein	FE556006.1	1	569	GATTTGCTGCAGGGCCAGTCGGTGCCAGGG
hypothetical protein	XM_002602515.1	1	632	GCAAAAACATAAAAGTCTGGGAAGGTTCAC
hypothetical protein	BW842917.1	1	576	GGAAAAAAATAAAAATGGCGGCGTACAAGT
transcribed locus	BW738825.1	1	316	GGTTACTGTGTGAATTGACGGCCATTTTGT
transcribed locus	BW696487.1	2	800	GTACAAAACGAACGAACGCACCTCAGAAAA
hypothetical protein	XM_002595158.1	2	473	GTACATCGGACAGAAAACCGCAGCTGCTAA
hypothetical	XM_002607896.1	1	558	GTAGAAAAATTCACCAGGAGCTCGTGTAGT

protein				
hypothetical protein	XM_002612187.1	3	447	GTCCACCAATATAGCGATCGTGGCATCAAG
hypothetical protein	XM_002608680.1	1	518	GTCTGCTGCACAACTGATCAGTGCGGCTTC
hypothetical protein	XM_002588769.1	5	784	GTTCCACACCTGACGTCCATTAAAAAAACG
histone H2A.x	CF918283.1	1	621	GTTCTTCATCTCCTGTGGATCGGGGCCAGA
hypothetical protein	XM_002590267.1	1	445	GTTGTGCGCTACTTCTGAAGTCCGGACGAC
hypothetical protein	XM_002607060.1	2	574	GTTGTGGTACCACAGAAAAAAAAAACAGTA
iron-sulfur cluster assembly enzyme ISCU, mitochondrial isoform ISCU2 precursor	XM_002611741.1	3	623	GTTGTTATACATATATCGTGGTGCGAAGGT
hypothetical protein	FE587424.1	1	531	TAAAAACAAGATGGCTGCGCCCTTGGAAGC
26S proteasome non-ATPase regulatory subunit 4	XM_002608276.1	2	858	TAAAAGAGTAACAGACGTCAAGATGGTGCT
hypothetical protein	BI380314.1	1	657	TAAATAAACACAGTTACCCAGTAAACCTAT
T-complex protein 1 subunit epsilon	FE594247.1	1	87	TACCACAGGAACCTGCAGGGCAACACATCT
hypothetical protein	BW701527.1	1	103	TACTAGTAGCTCTCGACTCGCGAAGTCTCA
hypothetical protein	XM_002585721.1	2	506	TACTGTCGTGACATCTAACGTTACATCTAC
zinc finger protein 389	FE564038.1	1	547	TAGTTCACCTCTGACCCCTTGCTAGGACGA
transcribed locus	BW796686.1	2	530	TCAGGTTCGGAAAAACCCCTACCAGAGCAA
histone H2A.V isoform 1	XM_002610290.1	3	381	TCAGTGGCGTTGCTTTGTTGTGCAGAGAAG
growth arrest-specific protein 1-like	FE543238.1	1	730	TCATAAACACAGCCGCATCCCCATCGAGAC
chromobox protein homolog 1	BW733933.1	4	430	TCATTGCTACACACTGAAAAGCAGTGCTGT
hypothetical protein	BW711318.1	3	642	TCCACTCGTACATCAAAACGGTGAGAGTTA
hypothetical protein	BW696695.1	1	523	TCCTTTGCGAAATACCATTGCCGACCGCCA
cytochrome b-c1 complex subunit 7 isoform 1	XM_002607528.1	1	122	TCTTTCCGGCAAAAATGGCGTCGCGAGTCG
transcribed locus	FE569161.1	28	445	TTAAAGTTTTTCTGGAGGAGTTCTGCAGAG
transcribed	BW704432.1	3	472	TTAATTGTTTTGCCTGAAGACATCCTGGTC

locus				
hypothetical protein	XM_002601085.1	2	502	TTACTTCCTCCCGACATCTCTAGACGAAAA
histone H3.3	XM_002595577.1	2	754	TTCCTGTCTGGCCGGAGAATTTGACTTCAA
similar to TBP-associated factor 15 isoform 1 isoform 2	XM_002586822.1	11	587	TTCGACTCTGGACATTTTCTGTGTAAGAAC
pre-rRNA-processing protein TSR2 homolog	XM_002601692.1	1	545	TTCGTCAAAAACCAAAAATGGCTGCGCCCA
hypothetical protein	XM_002600382.1	1	517	TTCTCCTCTACACCTTCGGCACAGACGGGT
mitochondrial 2-oxodicarboxylate carrier isoform 2	BW846189.1	4	554	TTGCTAAATACGTCATAGCGCGGTGGGATC
RNA polymerase-associated protein RTF1 homolog	XM_002589161.1	4	630	TTGTGGGAAAGAAGTGGACGATATGAAGAT
hypothetical protein	XM_002604528.1	1	691	TTTATTTTCCGTCAAAAACAAAAACGGCTC
transcribed locus	BW743122.1	1	381	TTTCGGTTCAGGGCACCCGGTCTAGCACTA
hypothetical protein	XM_002587537.1	8	605	TTTCTGGAAAGCGGTGCAACTCCTGGTGGT
cold-inducible RNA-binding protein	FE541241.1	8	75	TTTGCATACTGCCTTCAGATTGTGTTCATC
serine/arginine-rich splicing factor 2	FE571589.1	3	344	TTTGGAGGAGAAGGCGCGCCATAGTGAAGA
matches starting from 1st base	BW893234.1	1	30	TTTTCCTCATGACGAGATAGTGTCGGCTGT
transcribed locus	BI382259.1	1	622	TTTTCCTGCTACGCCACAGCGCATCTTGTG
cytochrome b5 isoform 1	XM_002607365.1	1	512	TTTTTCAGACAAAAAAGTGAGGTCACCCGA
polyubiquitin-B precursor	FE559758.1	1	132	TTTTTCCAGACAAGAAAGAGATTCGATTGA
hypothetical protein	BI388134.1	1	412	TTTTTCGCACCTGCCACTCCAAGGGTAAGG
cytochrome c oxidase subunit 4 isoform 1, mitochondrial precursor	BW699342.1	2	392	TTTTTTCCGGGACTGTGAAAGGAGCAGAAT
cytochrome c oxidase subunit 6B1	FE558478.1	4	453	TTTTTTTGATCTCCAAAGAGGAGAAGTGAC

*****60S ribosomal protein L32, one its four reads did not contain the 5’ TOP sequence, instead sequence was:

TCTTGGATCCTGGAAAAGGGATCTGAAACACCATCATGGTTGTGCGTCCGCTGAGGAAG
CCGAAGATCGTGAAGAAGCGGGTGAAGAAAATTCATCCGCCATCAGAGCGACAGATATG
ACAAACTAAAGCAAAACTGGCGTAAGCCCAAGGGTATCGACAACCGTGTGCGTAGGCGC
TTCAAGGGCCAGTACCTCATGCCCAACATCGGTTATGGGAGTGCCAAGAAGACAAAGCA
CCTTCTGCCCTCCGGCTTCAAGAAATTCCTCGTTCACAATGTCACTGACTTGGAGGTCT
TGTTGATGCAGAACCCGTACCTTCTGTGCCGAGATCGCTCACAACGTGTCGTCACGCAA
GAGGAAACTCATCGTGGAGCGCGCCCAGCAGCTTGGCCATTCCAAGGTTCCACCAACCC
AACCGAACCNCCNCCNCCGGCCTTCCGGCCGGAAGGAAGGAA

******60S ribosomal protein L28 isoform 2, three of its fifteen reads did not contain the 5’ TOP sequence, instead all three reads (545, 589 and 71 bases respectively) began:

GAGGATCTGCAGTGGATGATCATCAGGAACAACTCCTCCTTCTTGCTGAAGAGGAACAA
GCAGAGCTT…

APPENDIX VIII – NON-ALIGNING 5'-ENDS OF TRIMMED OLIGO-CAP 5'-RACE READS

In alphabetical order of non-aligning sequence.

BRANCHIOSTOMA

5'-end sequence not included in genomic BLAT alignment	**Gene #**
CAAACTCTG	170
CAACA	181
CAATCGCTGGTCCCTGTCTCCACTGAAGGTTTA	94
CACA	58
CACA	33
CACA	188
CAGCA	32
CCA	187
CCAGCGTCCACAGAGCACATGCGAGAC	14
CCGT	73
CGACACTGACAGGGAACTTGAAAAGGAGTTAGAAA	129
CGAGAAAGCGAGATG	6
CTA	90
CTA	62
CTATTGCCGACCGCCATCTTGCAGCCAGGACAGAGCCACCGA	148
CTC	62
CTC	62
CTC	62
CTC	151
CTC	114
CTC	62
CTC	114
CTC	62
CTC	62
CTCA	111
CTCT	55
CTCT	62
CTCT	55
CTCT	114
CTCT	180
CTCT	121
CTCT	55
CTCT	78
CTCT	55
CTCTTCGCCATTTTCCA	5
CTCTTCGCCATTTTCCA	5
CTCTTCGCCATTTTCCA	5
CTCTTTCATTTCCGACG	177

CTCTTTCATTTTCCGGCG	193
CTCTTTCATTTTTCCGACA	193
CTCTTTCCTGAAACCCCAAAAATCCAAGATGGGGTAAAGC	166
CTCTTTCCTTGCCA	77
CTCTTTCCTTGCCACCATTTTTCA	77
CTCTTTTCATTTTTCCGACA	193
CTT	151
CTT	99
CTT	95
CTT	185
CTT	95
CTT	95
CTTC	2
CTTC	135
CTTCCG	2
CTTCCGAAGAGGCGAAAAAT	85
CTTCGAGTG	125
CTTCGAGTG	125
CTTCTTACTTGCGCCCAGCCAAGTCACAT	158
CTTT	80
CTTT	55
CTTTCTGGCCGCCATTACT	11
CTTTCTGGCCGCCATTACT	11
CTTTTT	101
GAA	78
GAA	78
GAAA	10
GAAAGAA	27
GAAAGAA	26
GAAATAAACT	167
GAC	143
GACA	130
GACACT	197
GACGCCATTTTTTGTAGAGAGAGATTTGATGCAAGGAATTAGTCA	165
GCACACCGAACACACACCTCCACACCGAAGTCCATGGAGTCGACTCGCGAGTC CAGCCCAGACACCCACAGGGACGTACGT	141
GGAA	93
GGCGAGTCG	41
GGCTGACCCC	175
GTAA	11
GTTAAATA	76
GTTGA	7
GTTGA	7
TAA	73
TAA	20
TAAACT	167

TAAACT	167
TAAACT	150
TAAACT	167
TAACT	167
TAACT	167
TAACT	167
TAACT	150
TAACT	167
TAACT	167
TAACT	167
TAACT	167
TAACT	150
TAACT	167
TAACT	150
TAACT	150
TAACT	167
TAACT	167
TAACT	150
TACAC	22
TACCAG	133
TACCAGACCATCTAAATCTACACGGGCTGTTCTCGGCACTGAC	90
TAGGCTTCGTTGCAGAACTTTG	34
TATA	141
TATA	140
TATAAACT	167
TATAACT	167
TATAACT	150
TATAACT	167
TATAACT	167
TATAACT	150
TATAACT	167
TATAACT	167
TATAACT	167
TATAACT	167
TATAACT	167
TATAACT	150
TCA	98
TCA	106
TCCA	177
TGAA	164
TGACCAGCAGATCGGACAAAA	191
TGTA	135
TGTA	135
TGTGAA	50
TGTTC	106

TTA	190
TTAACT	150
TTACTTCCTCCCGACATCTCTAGACGAAAGACAAGACGGACAAAGCAGGAACA	89
TTCA	186
TTCAAACCG	183
TTCAAACCG	183
TTCAAACCG	183
TTCAAACCG	183
TTCCAAGTCTGACTACGAGCGGAGTCGTGCCCTATCAACCAAACAAAAA	103
TTCGACTCTGGACATTTTTCTGTGTAAGAAAC	45
TTCTT	192
TTGCCGACCGCCATCTTGCAGCCAGGACAGAGCCACCGA	148
TTT	86
TTT	136
TTTA	47
TTTT	86
TTTT	165
TTTTA	101
TTTTACATACTG	80
TTTTACTTGCGCCAGCCAAAGTCACAT	185
TTTTACTTGCGCCAGCCCAAGTCACAT	185
TTTTGGTCCGAGGCTTTGAT	40
TTTTTTTCCTGAAACCCCAAAAAATCCAAGATG	166

SACCGLOSSUS

5'-end sequence not included in genomic BLAT alignment	**Gene #**
CAAAAGATGGCGTC	33
CATCTAAATCAGCATATTTGGGTGGAACACAGACCGACACATCAACAGGCAAC ATGGCTGAGGCAGAACAACTAACCGAGGAACAGATTGCT	31
CATCTAATCAGCATATTTGGGTGGAACACAGACCGACACATCAACAGGCAACA TGGCTGAGGCAGAACAACTAACCGAGGAACAGATTGCT	31
CATCTAATCAGCATATTTGGGTGGAACACAGACCGACACATCAACAGGCAACA TGGCTGAGGCAGAACAACTAACCGAGGAACAGATTGCT	31
CATCTAATCAGCATATTTGGGTGGAACACAGACCGACACATCAACAGGCAACA TGGCTGAGGCAGAACAACTAACCGAGGAACAGATTGTT	31
CCT	105
CGACGACTACGTACGACTAGTACGTACGTAGAGTAGAAGAGACGAGACGACG ACGACGACGACGACGACGTACGTACGTACGACGACGTAGAAGACGTAGTACG ACGACGACGACGACGACGACGACGACGACGA	86
CTTCGGGACTACAG	35
CTTT	40
CTTTTCTCGTAGCAGTCAAAAAT	78
CTTTTCTCGTAGCAGTCAAAAAT	78
CTTTTCTCGTAGCAGTCAAAAAT	78
CTTTTCTCGTAGCAGTCAAAAAT	78
CTTTTCTCGTAGCAGTTAAAAT	78

CTTTTCTCGTAGCAGTTAAAAT	78
CTTTTTAGCTGTACTGATAAGCACAGCCCGTCATGCCT	53
GAGAGAAAGAACACTCGAAGAACCCGCAATATCGTCACAATGCTGATCAAAG TTAAGACACTCACGGGGAAGG	48
GAGAGAAAGAACACTCGAAGAACCCGCAATATCGTCACAATGCTGATCAAAG TTAAGACACTCACGGGGAAGG	48
GAGAGAAAGAACACTCGAAGAACCCGCAATATCGTCACAATGCTGATCAAAG TTAAGACACTCACGGGGAGGG	48
GCACTGACATGGACTGAGGAGTAGAA	149
GTA	47
GTA	47
GTA	47
TAGTA	32
TAGTAAAATTTA	32
TAGTAAATTTA	32
TAGTAAATTTATCAA	32

STRONGYLOCENTROTUS

5'-end sequence not included in genomic BLAT alignment	**Gene #**
CAGATAA	8
CAGATATCATCAATCATCAGCCACAAACATACCGTATGTGATCTTGACTAGAT CTACTTGTATTTAAGAAAAGGGAAGAATATGTCAACTTCCTTGGTGCGGAAAA GGCCTGGAACTGTTCGCAGATGAATCT	56
CATTCCCCTCTCCTACTCCTACTCCTACTCCTACTCCTACTACTACTACTCATAT TCCCTCTCCTACTCCTACTCCTACTCCTACTCCTACTCCT	1
CATTTCTCGCAAAACCAGCAGCATTCATCGTCGATTTGCTCCACTGTGAAAACA TCACTACTGTATGTTGTCCCGATTTGTATTTATTTAATACACTTCAGGAAATAA TTCAAAAATGGCTGAATATTCGGCACGTACGGATAAGACCGTGTCAGACAGAT ATATGACCCCTTGACACGGATGGCCATTGTCAGGTCATGTACATCTGGGTTGAT GGCTCGGGAGAGAACCTAAGAGCCAAAACGAAAACGATGGATACCGTACCTG AAAAAACCCTGA	53
CCTTTTTCTCTGATCCGTCGAGCATATAAAAATCCTGTCAAGATGCAAAAACG AAGGAGGAGAATACGTGGATATGTACACTCCACGAAAGT	68
CGTACGCGAACACAACGCTCAGGCGAGACATTTATTTTGCGCAAAAATCGTCT AACTTATTTGAGTAATAAACCATTTAAGCATGGATGAGGACGATAGAAAAAT	28
CTCT	13
CTCTATCTATCAAAACATGGC	16
CTCTTAAAAGTCTGCTGCCCTCCCGAAGAT	23
CTCTTAAAAGTCTGCTGCTCTCCGAAGAT	23
CTCTTAAACTATCCCTAAACTAAC	55
CTCTTAAACTATCCCTAACTAAC	55
CTCTTACACCTCTAGCCGTTCAAAAATG	26
CTCTTGACGTGAGATATCTCTCAAAAATGGTGAACGTTCCAAAGAACCGCCGC ACATACTGCAAGAAGTGCAAGAAGCACATGAACCACAAGGTCACCCAGTACA AGGCTGGTAAAAGCTTCACTCTACGCCCAGGGTAAGAGGCGTTACGACAGGA AGC	45
CTCTTGACGTGAGATATCTCTCAAAAATGGTGAACGTTCCAAAGAACCGCCGC ACATACTGCAAGAAGTGCAAGAAGCACATGAACCACAAGGTCACCCAGTACA AGGCTGGTAAAGCTTCACTCTACACCCAGGGTAAGAGGCGTTACGACAGGAA	11

G	
CTCTTTCCAAAAATATCAGTAGAGGGGTCTAAAAACT	51
CTCTTTCCAAAAATATCAGTAGAGGGTCTAAAAACT	51
CTCTTTCCATCCAAAAAATGTCGAAGCGAGGACGTGGAGGTTCATCTGGATCTAAAATTCCGTATCTCACTGGGTTTTGCCAGTGGGAGCGGTCATCAACTGTGCAGACAATAC	31
CTTC	9
CTTTTTTCCCTGATCCGTCGAGCATATAAAAATCCTGCCAAGATGCAAAAACGAAGGAGGAGAATACGTGGATATGTACACTCCACGAAAGT	68
CTTTTTTCTCTGATCCGTCGAGCATATAAAAAATCCTGCCAAGATGCAAAAACGAAGGAGGAGAATACGTGGATATGTACACTCCACGAAAAGT	68
CTTTTTTTCGTTGAAACTCAAACAGACAGCA	50
GAA	12
GAAGAAGGCCGAACCTGGTAGTTTAGATGATCTTCTACTGCAGAAACTCAAAAGGGAGTAAAAGGGTTAGAAAAGATTGGCAGATTCAAAAGAAGAAAAAAGAAGAAAAAATAAGAAGATTGGCCGGGATACAGATCAGTTGTCAAAAGAGAAAACAAGAAGAGAAAAAGAAAGGCATCACCATAGCGAGAGTGATTCAGATTCTACAGAACAATCAGAGAGT	7
GATTCCTGCTA	38
GCTA	61
GGCGG	20
GTATCCATTGACGAATCTTGAAAAAACGCTGGCGTCTTCTGCGGTAGAAGTCAACAAAAAAACTCTACAAAAAAT	52
TAAAAGAA	12
TATATCAAATGCAGATCGACTATTGAAGTAGACTTTCTAGGCCTAAAAGATTGTAAC	3
TCATT	66
TTAAAAGAA	12
TTATATCT	30
TTTTTTACGTTTCTAGTCTGTGTCTTTGTGTGATCGTGCGCTGATTCTTGATTGCCGGAGATCGCGAGAAAAAACTT	37