Computational analysis of spliced leader trans-splicing in the regenerative flatworm *Macrostomum lignano* reveals its prevalence in conserved and stem cell related genes

K.V. Ustyantsev, E.V. Berezikov ✉

Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ eberez@bionet.nsc.ru

Abstract. In eukaryotes, trans-splicing is a process of nuclear pre-mRNA maturation where two different RNA molecules are joined together by the spliceosomal machinery utilizing mechanisms similar to cis-splicing. In diverse taxa of lower eukaryotes, spliced leader (SL) trans-splicing is the most frequent type of trans-splicing, when the same sequence derived from short small nuclear RNA molecules, called SL RNAs, is attached to the 5′ ends of different non-processed pre-mRNAs. One of the functions of SL trans-splicing is processing polycistronic pre-mRNA molecules transcribed from operons, when several genes are transcribed as one pre-mRNA molecule. However, only a fraction of trans-spliced genes reside in operons, suggesting that SL trans-splicing must also have some other, less understood functions. Regenerative flatworms are informative model organisms which hold the keys to understand the mechanism of stem cell regulation and specialization during regeneration and homeostasis. Their ability to regenerate is fueled by the division and differentiation of the adult somatic stem cell population called neoblasts. *Macrostomum lignano* is a flatworm model organism where substantial technological advances have been achieved in recent years, including the development of transgenesis. Although a large fraction of genes in *M. lignano* were estimated to be SL trans-spliced, SL trans-splicing was not studied in detail in *M. lignano* before. Here, we performed the first comprehensive study of SL trans-splicing in *M. lignano*. By reanalyzing the existing genome and transcriptome data of *M. lignano*, we estimate that 30 % of its genes are SL trans-spliced, 15 % are organized in operons, and almost 40 % are both SL trans-spliced and in operons. We annotated and characterized the sequence of SL RNA and characterized conserved cis- and SL trans-splicing motifs. Finally, we found that a majority of SL trans-spliced genes are evolutionarily conserved and significantly over-represented in neoblast-specific genes. Our findings suggest an important role of SL trans-splicing in the regulation and maintenance of neoblasts in *M. lignano*.

Key words: flatworms; regeneration; splicing; trans-splicing; neoblasts; spliced leader; *Macrostomum lignano*.

Биоинформационный анализ сплайс-лидерного транс-сплайсинга у регенерирующего плоского червя *Macrostomum lignano* показал его преобладание среди консервативных генов и генов стволовых клеток

К.В. Устьянцев, Е.В. Березиков ✉

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ eberez@bionet.nsc.ru

Аннотация. Транс-сплайсинг у эукариот – это процесс созревания ядерных пре-мРНК, когда две различные молекулы РНК соединяются с помощью структур сплайсосомы по механизму, схожему с цис-сплайсингом. У различных таксонов низших эукариот наиболее распространенный тип транс-сплайсинга – сплайс-лидерный (СЛ) транс-сплайсинг, при котором одинаковая последовательность, происходящая от коротких малых ядерных РНК молекул, называемых СЛ РНК, присоединяется к 5′-концам различных непроцессированных пре-мРНК. Одна из функций СЛ транс-сплайсинга состоит в процессировании полицистронных молекул пре-мРНК, транскрибируемых с оперонов, когда транскрипция нескольких генов осуществляется как одна молекула пре-мРНК. Однако лишь часть генов, подвергающихся транс-сплайсингу, содержится в оперонах, что говорит о том, что у СЛ транс-сплайсинга должны быть и другие, менее изученные, функции. Регенерирующие плоские черви являются информативными модельными организмами, хранящими ключи к пониманию механизмов регуляции стволовых клеток и их дифференцировки во время регенерации и при гомеостазе. Их способность к регенерации – следствие деления и дифференцировки соматических стволовых клеток, называемых необластами, которые присутствуют у взрослых особей. *Macrostomum lignano* – модельный плоский червь, в исследованиях на котором в

последние годы достигнут существенный технологический прогресс, включая разработку метода трансгенеза. Сплайс-лидерный транс-сплайсинг ранее не был детально изучен у *M. lignano*, хотя известно, что значительная часть генов *M. lignano* подвергается этому типу транс-сплайсинга. В настоящей работе мы осуществили первое обширное исследование СЛ транс-сплайсинга у *M. lignano*. Повторно проанализировав геномные и транскриптомные данные *M. lignano*, мы оцениваем, что 30 % его генов подвергаются СЛ транс-сплайсингу, 15 % расположены в оперонах, а почти 40 % находятся в оперонах и проходят через СЛ транс-сплайсинг. Мы провели аннотацию и охарактеризовали последовательность СЛ РНК и консервативных мотивов цис- и транс-сплайсинга. Обнаружено, что большинство генов, подвергающихся СЛ транс-сплайсингу, эволюционно консервативны и значительно перепредставлены в генах, специфичных для необластов. Наши результаты предполагают важную роль СЛ транс-сплайсинга в регуляции функционирования необластов у *M. lignano*.

Ключевые слова: плоские черви; регенерация; сплайсинг; транс-сплайсинг; необласты; сплайс-лидер; *Macrostomum lignano*.

Introduction

Before being used as templates for protein production, majority of RNA molecules transcribed in the nucleus (pre-mRNA) undergo three major modifications to become mature and fully functional mRNA. This is called RNA processing and it involves capping of the 5′ end, polyadenylation of the 3′ end, and splicing. Two types of splicing are distinguished – cis- and trans-splicing. During cis-splicing all the processing happens with the same pre-mRNA molecule, resulting in the removal of introns and merging of its exons. During trans-splicing, on the other hand, two different pre-mRNA molecules expressed from distinct genomic loci are joint into a new chimeric trans-spliced mRNA (Lasda, Blumenthal, 2011).

Trans-splicing was originally discovered in trypanosomes (Euglenozoa), where it was found that a short 39 bp leader sequence was post-transcriptionally attached to the 5′ ends of variant surface glycoproteins pre-mRNA (Boothroyd, Cross, 1982). Later, 5′ end addition of a 22 bp spliced leader (SL) was also observed in *Caenorhabitis elegans* mRNA of actin gene and some other genes (Krause, Hirsh, 1987). Now this process is well known as SL trans-splicing. A distinct feature of SL trans-splicing is that all such processed transcripts have the same short SL sequence, or its variant, at their 5′ ends. The SL sequence is derived from an exon of a non-coding small nuclear RNA molecule called SL RNA, which is ~100 nt in length and has 2,2,7-trimethylguanosine cap at its 5′ end instead of 7-methylguanosine cap, which is found in non-trans-spliced mRNAs (Liou, Blumenthal, 1990; Lasda, Blumenthal, 2011). SL RNAs have a splicing donor site at the exon 3′ end, while SL trans-spliced pre-mRNAs have a splicing acceptor site at the 5′ end of their first exon. SL trans-splicing results in removal of the 5′ non-exon pre-mRNA part called outron (Lasda, Blumenthal, 2011). It is experimentally shown that the only requirement for a gene to be predominantly SL trans-spliced is an acceptor splicing site close to the 5′ end of the first exon that is not complemented by a donor splicing site upstream in cis (Conrad et al., 1993). Thus comes another important feature of SL trans-splicing, namely that it allows formation and resolving of operons – adjacent genes transcribed as a single pre-mRNA from the same promoter region (Blumenthal, Gleason, 2003). However, apart from a clear function in polycistronic transcripts resolution, the function of SL trans-splicing for monocystronic transcripts is still in debate (Danks et al., 2015). It is hypothesized that the function may be in equalization of 5′ UTRs in length and their clearance from out-of-frame AUG start codons, while at the same time allowing less restricted evolution of 5′ upstream regulatory sequences, and in additional control of translation (Hastings, 2005; Danks, Thompson, 2015). So far, SL trans-splicing was found in several clades of eukaryotes: dinoflagellates, euglenozoans, cnidarians, flatworms, nematodes, and ascidians (Lei et al., 2016). SL trans-splicing is most prominent in trypanosomes (100 % genes are trans-spliced) and in nematodes (70 % genes are trans-spliced) (Allen et al., 2011; Lei et al., 2016).

Regenerative flatworms are informative models to understand the mechanism of stem cell regulation and specialization during regeneration and homeostasis. Their ability to regenerate is driven by the division and differentiation of the adult somatic stem cell population called neoblasts (Wagner et al., 2011; Mouton et al., 2018). *Macrostomum lignano* is the only flatworm species for which a method for stable transgenesis is available so far. The worm also has a number of features allowing for efficient cell lineage tracing and phenotype screening, which makes *M. lignano* an attractive model to study a wide range of biological processes (Grudniewska et al., 2016; Wudarski et al., 2017, 2019, 2020). Well-annotated *M. lignano* genome and transcriptome assemblies were recently published (Wudarski et al., 2017; Grudniewska et al., 2018). It was estimated that almost 21 % of its genes are SL trans-spliced to the same 35 bp SL sequence (Grudniewska et al., 2018). However, trans-splicing was not studied in details in *M. lignano*, and its impact on the genome functioning and maintenance is still unknown. Here, we present the first comprehensive study of SL trans-splicing in *M. lignano* and show that it is strongly connected with genes specific for the neoblasts of the worm.

Materials and methods

Data. The published *M. lignano* genome Mlig_3_7 (Wudarski et al., 2017) and transcriptome Mlig_RNA_3_7_DV1_v3 (Grudniewska et al., 2018) assemblies and the corresponding annotation tracks were obtained from (http://gb.macgenome.org/downloads/Mlig_3_7/).

Genome deduplication. Mlig_3_7 genome assembly was deduplicated using purge_dups software (v1.0.1) with default settings (Guan et al., 2020) and utilizing published PacBio genome sequencing data (Wasik et al., 2015) for the calculation of contig coverages. Contig names from the deduplicated genome assembly were used to extract respective gene annotations from the full Mlig_3_7 genome annotation.

Motif discovery and SL RNA annotation. Presence of the SL sequence at the 5′ end of the *M. lignano* transcripts was established in the previous studies (Wasik et al., 2015; Grudniewska et al., 2016). For the annotation of trans-spliced genes, SL-containing RNA-seq reads were mapped to the

К.В. Устьянцев
Е.В. Березиков

Стволовые клетки и транс-сплайсинг
у регенерирующего плоского червя *Macrostomum lignano*

2021
25•1

Mlig_3_7 genome assembly and the presence of such reads at the beginning of transcripts was used as an evidence of SL trans-splicing (Wudarski et al., 2017; Grudniewska et al., 2018). Therefore, all the SL trans-spliced transcripts have the corresponding annotation in the Mlig_3_7 genome assembly, and the sequences upstream of their first exon were considered as outrons. Using the deduplicated annotation track of gene coordinates, we retrieved nucleotide sequences of genomic regions corresponding to exon-intron and exon-outron (for the trans-spliced genes) junctions with 50 bp flanks in both directions. All the sequences were converted to forward orientation and split into three groups corresponding to cis-donor, cis-acceptor, and trans-spliced acceptor sites. The sequences then were analysed for the presence of enriched motif using a stand-alone version of the DREME tool (Bailey, 2011).

To determine the SL RNA gene sequence in the genome assembly, we used the 35 bp *M. lignano* SL sequence (CGG TCTCTTACTGCGAAGACTCAATTTA TTGCATG) as a seed for a BLASTn (Altschul et al., 1990) search requiring only 100 % matching hits. Next, we manually investigated genomic sequences surrounding the BLAST hits by matching the SL sequence track in the genome browser to the expected size of SL RNA (~100 bp). The corresponding sequences were then checked for folding into secondary structure canonical for SL RNA folding using Mfold web server (Zuker, 2003), and conserved motifs were then manually identified.

Prediction of operons. Intergenic distances were retrieved from the deduplicated genome annotation file. We only considered distances between immediately adjacent transcripts with the same transcriptional orientation and not interrupted by transcripts in opposite orientation. Distances were split into three categories: between SL trans-spliced genes, between a non-SL trans-spliced gene and an SL trans-spliced gene, and between non-SL trans-spliced genes. To adjust for repetitive element insertions, we retrieved the corresponding coordinates from the genome browser RepeatMasker and TRF tracks (http://gb.macgenome.org/) and subtracted them from the previously identified intergenic distances. Distribution of the distances was visualized as density plots using ggplot2 library in R.

After the analysis of the graphical data of the distances distributions, we selected the threshold value of 1000 bp, below which a pair of adjacent and SL trans-spliced genes were considered as belonging to the same operon. The same applies if the first gene is non-trans-spliced, but the second is SL trans-spliced. The distributions of lengths of operons of various sizes was visualized as violin plots using ggplot2 library in R.

Estimation of gene conservation. Gene annotation and data classifying genes as being specific to neoblasts or germline were retrieved from the previous study (Grudniewska et al., 2018). A gene was considered to be conserved if it has an open reading frame with a detectable homology to a human gene, which is indicated in its annotation, and non-conserved if lacking the homology to human, but has a predicted open reading frame with homology to proteins in other organisms. Otherwise, a gene was considered non-coding.

Results
Deduplication of genome assembly.
The published Mlig_3_7 genome assembly is based on the sequencing data from DV1 *M. lignano* line. This line has a $2n = 10$ karyotype (four large and six small chromosomes) and was demonstrated to have undergone a duplication of its large chromosome (Zadesenets et al., 2017), while the karyotype of the basal wild type population is $2n = 8$ (two large and six small chromosomes) (Wudarski et al., 2017). The size of Mlig_3_7 assembly is 764 Mb, which corresponds to the experimental measurement of the genome size in the DV1 line (Wudarski et al., 2017), and the assembly contains the duplicated large chromosome sequences. To avoid gene overcounting due to the presence of these duplicated sequences in the Mlig_3_7 assembly, we removed the most redundant scaffolds by deduplicating Mlig_3_7 assembly using purge_dups software (Guan et al., 2020). This resulted in approximately 46 % drop in the number of scaffolds (from 5270 to 2841) and decreased the genome size to 580 Mb, which is close to the genome size measurements for the NL10 line of *M. lignano*, which does not have the chromosomal duplication (Wudarski et al., 2017). Next, we removed the records from transcriptome annotation which corresponded to the redundant scaffolds.

Motif discovery and SL RNA gene mapping. Investigation of the deduplicated part of the transcriptome shows that a significant fraction of genes, 21 754 out of 71 499 (30 %), are SL trans-spliced in *M. lignano.* This means that they all have the same 35 bp SL sequence (CGGTCTCTTACTGCG AAGACTCAATTTA TTGCATG) at the 5′ end of their processed transcripts (Wudarski et al., 2017; Grudniewska et al., 2018). Despite this, SL trans-splicing was not characterized in more detail in *M. lignano*. First, we retrieved genomic DNA sequences near the cis-splicing and SL trans-splicing exon-intron/exon-outron junction sites and checked if they are enriched for some motifs using DREME (Fig. 1, *a*) (Bailey, 2011). In total, we obtained 187 627 regions around 5′ donor and 3′ acceptor cis-splicing sites and 21 754 regions around SL trans-splicing sites. The first most enriched motifs near cis-splicing 5′ donor and 3′ acceptor sites were GT[G/A]AG (found in 122 399 regions, p-value: $8.8\mathrm{e}^{-23468}$) and CAG (found in 112 174 regions, p-value: $1.7\mathrm{e}^{-12459}$), respectively, corresponding to canonical cis-splicing motifs. A motif [T/C] TNCAG (found in 9551 regions, p-value: $1.3\mathrm{e}^{-1631}$) was the top enriched motif near SL trans-splicing 3′ acceptor sites. All the motifs were positioned right at the exon-intron/exon-outron junctions of the corresponding sites (see Fig. 1, *a*).

Next, to confirm the presence of the SL RNA gene in the genome assembly, we analysed the secondary structure of the previously published sequence of *M. lignano* SL RNA from the ML2 version of the genome (Wasik et al., 2015). However, we found that the reported sequence was erroneously assigned as SL RNA, since it clearly maps to the 5′ end of an SL trans-spliced protein-coding gene (Mlig013257.g1, scaf577:45 663-48 770) in the Mlig_3_7 assembly, and also does not fold into canonical structure with three hairpin loops (data not shown) (Xie, Hirsh, 1998). Therefore, we decided to identify the actual SL RNA gene in the newer Mlig_3_7 assembly. Using SL sequence as a seed for the genomic BLASTn search, we mapped a 109 bp sequence, which is repeated eight times in the deduplicated genome and has the canonical SL RNA secondary structure predicted by Mfold web server (see Fig. 1, *b*) (Zuker, 2003). Subsequent sequence analysis showed clear signatures of an SL RNA: the SL sequence
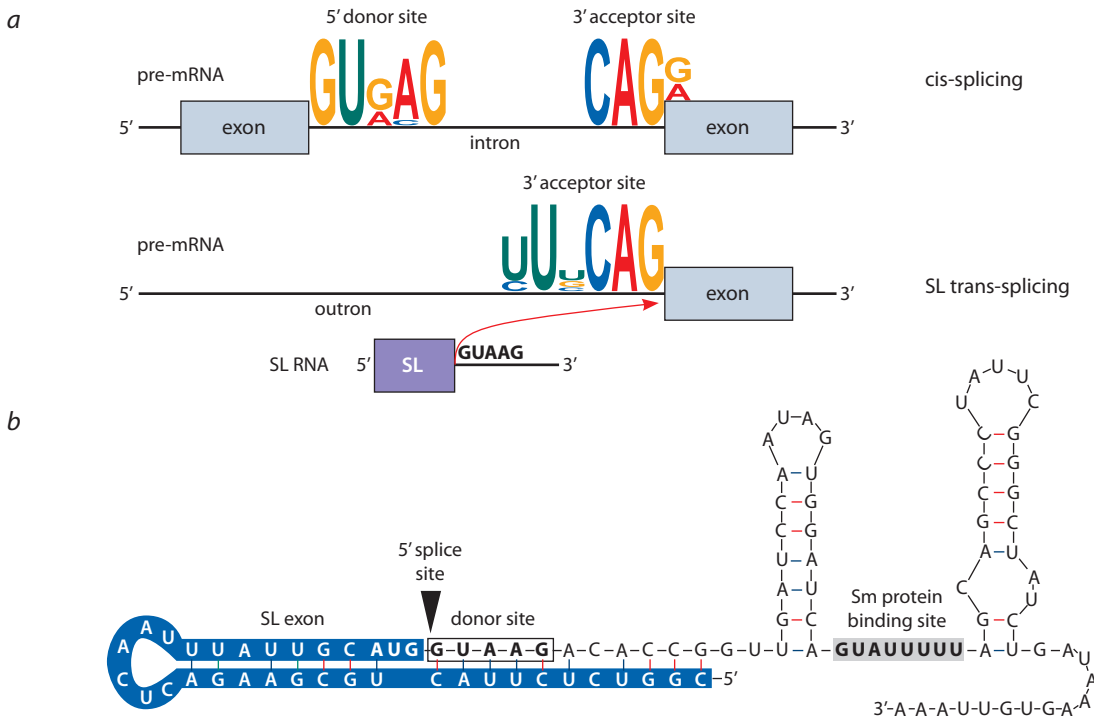
Fig. 1. Features of cis-splicing and SL trans-splicing in *M. lignano*.

a – conserved motifs enriched at the splicing junction sites in cis-spliced and SL trans-spliced genes; *b* – sequence and predicted secondary structure of *M. lignano* SL RNA gene.
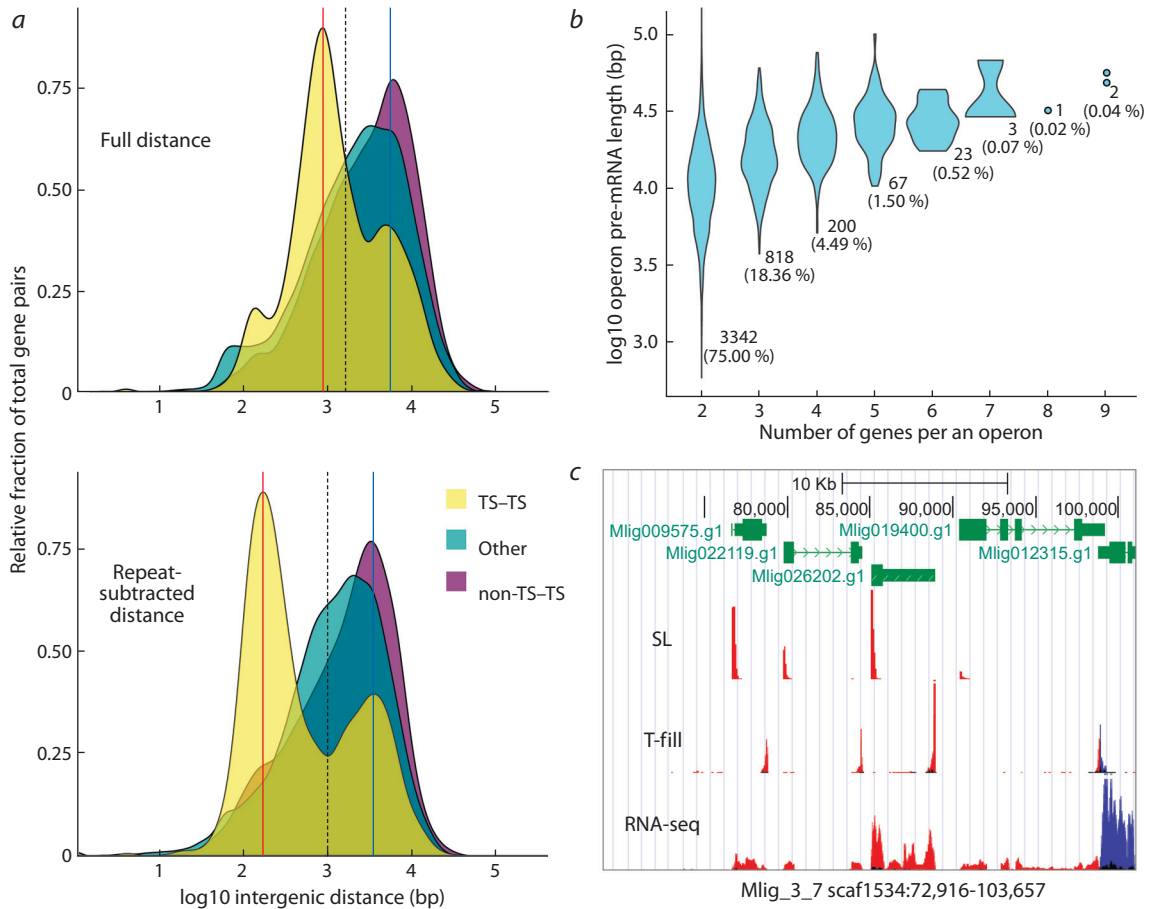
is at the 5′ end of the gene and forms the first hairpin loop, immediately after the SL sequence there is a clear 5′ donor splicing site (GTAAG), and between two other hairpin loops there is a motif similar to the binding site of Sm spliceosomal protein (see Fig. 1, *b*) (Ganot et al., 2004; Stover et al., 2006).

Operon analysis. The important feature of SL trans-splicing is that it allows for processing of long polycistronic pre-mRNA molecules expressed from a single promoter region in a way similar to prokaryote operons. In principle, genome-guided transcriptome assembly using RNA-seq data allows identification of such operons and their corresponding pre-mRNA sequences, which we previously annotated as transcriptional units (Wudarski et al., 2017; Grudniewska et al., 2018). However, it is not always possible to fully reconstruct an operon from RNA-seq data alone, since transcriptional units predicted from RNA-seq data tend to split in the repeat-rich intergenic regions of operons, where read coverage depends on both operon expression level and the frequency of repeats in the genome. Instead, to estimate what fraction of *M. lignano* genes are organized in operons based on their genomic organization, we first explored how intergenic distances between trans- and non-trans-spliced genes are distributed in *M. lignano* genome (Fig. 2, *a*). We found that distribution of distances between trans-spliced genes has multimodal distribution, while it is unimodal distribution for non-trans-spliced/trans-spliced and non-trans-spliced/non-trans-spliced intergenic distances (see Fig. 2, *a*). SL trans-splicing is an ancient evolutionary mechanism (Lei et al., 2016), which is mostly abundant in the genomes of simply organized organisms, which have low repetitive content and relatively small genomes (Gregory et al., 2007). We hypothesized that neutral

accumulation of repeats could have influenced the distances between genes in the operons. Interestingly, after we adjusted the intergenic distances by subtracting the fraction occupied by repetitive sequences (simple repeats and transposable elements), it had the most impact on the distances between trans-spliced genes, revealing a clear bimodal distribution with the most prominent peak at around 100 bp (see Fig. 2, *a*). This observation indicates that repeats have a substantial contribution to intergenic distances in operons. To classify genes as belonging to the same operon, we decided to use the repeat-adjusted distances with a threshold value of 1 Kb, which separates the two modes of the intergenic distances between trans-spliced genes (see Fig. 2, *a*).

Using these criteria for defining operons, we found that 10458 genes (approx. 15 % of all genes and 40 % of SL trans-spliced genes) can be assigned to operons, of which 1854 (18 %) start from a non-trans-spliced gene (see Fig. 2, *b*, Fig. 3). The vast majority of them are comprised of two and three genes (75 and 18 %), with the maximum operon size reaching nine genes (two operons) (see Fig. 2, *b*). An example of an operon defined in this way is provided in Fig. 2, *c*.

SL trans-splicing is enriched in evolutionary conserved and stem cell genes. We know from a previous study (Grudniewska et al., 2018) that evolutionary conserved protein-coding genes, which still have detectable homology between *M. lignano* and human, are enriched in somatic stem cells – neoblasts (85 % compared to overall 47 %) (see Fig. 3). On the contrary, only 38 % of germline-specific genes in *M. lignano* are conserved in human, suggesting their relatively recent appearance in evolution of flatworms. We investigated whether there is a correlation between gene conservation and

К.В. Устьянцев
Е.В. Березиков

Стволовые клетки и транс-сплайсинг
у регенерирующего плоского червя *Macrostomum lignano*

2021
25•1

Fig. 2. Identification and characteristics of genes in operons in *M. lignano* genome.

a – distribution of intergenic distances between various gene types. TS – SL trans-spliced, non-TS – non-SL trans-spliced. Red and blue vertical lines indicate modes of the distributions. Vertical black dashed line indicates distance threshold value selected to separate genes in operons; *b* – putative pre-mRNA length and abundance of different operon sizes; *c* – an example of an operon with four genes as depicted in the *M. lignano* genome browser (http://gb.macgenome.org). Genes are in green, with exons as blocks and introns as dashed lines. Non-protein-coding part of the exons are narrower. SL – RNA-seq reads mapped which contained the SL sequence at their 5' ends (trimmed). T-fill – RNA-seq reads mapped containing mRNA 3' poly-A ends. Reads mapped in forward orientation are in red, and the reversed reads are in blue.
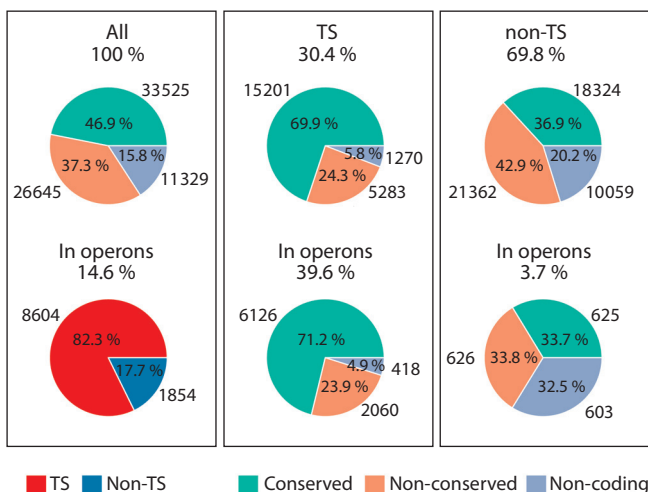
Fig. 3. Evolutionary conservation of *M. lignano* genes.

TS – SL trans-spliced; non-TS – non-SL trans-spliced; conserved – protein-coding genes with a homology to human; non-conserved – protein-coding genes lacking the homology to human; non-coding – genes do not code for a protein.

SL trans-splicing in *M. lignano*. We found that 69.9 % of the SL trans-spliced genes are conserved between *M. lignano* and human, while 24.3 % are not conserved and 5.8 % are non-coding (see Fig. 3). Trans-spliced genes that are located in operons have a very similar distribution of conserved, non-conserved and non-coding genes (see Fig. 3). In contrast, among non-trans-spliced genes only 36.9 % are conserved in human, while 42.9 % are non-conserved and 20.2 % are non-coding (see Fig. 3). Thus, SL trans-spliced genes are strongly enriched for conserved genes but there is no dependence on whether these genes are in operons or not.

Next, we calculated the fraction of SL trans-spliced genes among genes enriched in neoblasts (stem cells) and germ-line – the only proliferation capable cell types in the worm (Grudniewska et al., 2018). Intriguingly, 85 % of the stem cell genes (746) are SL trans-spliced, and almost 86 % (752) are conserved in human (see the Table), and 728 genes are both conserved in human and SL trans-spliced, which is 96.8 % of all the conserved genes in neoblasts. Given that out of 33 525 conserved genes present in the Mlig_3_7 genome annotation 15 201 (45.3 %) are trans-spliced (see Fig. 3), this represents

Summary of transcripts from *M. lignano* proliferation-capable cell types

Cell type	Total transcripts	Trans-spliced (%)	In operons (%)	Conserved in human (%)	Non-conserved in human (%)	Non-coding (%)
Neoblasts	878	746 (85.0)	343 (39.1)	752 (85.6)	19 (2.2)	107 (12.2)
Germline	1985	362 (18.2)	192 (9.7)	736 (37.1)	248 (12.5)	1001 (50.4)

a 2.13-fold enrichment for conserved SL trans-spliced genes among neoblast genes relative to the expected from the random distribution (p-value: 1.98e–7, chi-square test). On the contrary, only 18 % of the germline genes are SL trans-spliced and 37 % are conserved in human. Taken all together, this suggests that SL trans-splicing plays an important role in stem cell regulation in *M. lignano*.

Discussion

SL trans-splicing is widespread in diverse flatworm taxa, including both parasitic and free-living species (Zayas et al., 2005; Protasio et al., 2012; Wudarski et al., 2017; Ershov et al., 2019). However, most of the studies of SL trans-splicing were focused on nematodes and trypanosomes (Lasda, Blumenthal, 2011; Lei et al., 2016). Here, we performed the first study which focuses on SL trans-splicing in the free-living regenerative flatworm model *M. lignano*. By reanalyzing the available genome and transcriptome data, we found that 30 % of the worm genes are SL trans-spliced, and 15 % are estimated to be organized in operons (see Fig. 3). For a comparison, in *C. elegans* 70 % of genes are SL trans-spliced and 17 % are in operons, in ascidian chordate *Ciona intestinalis* it is 58 and 20 %, respectively, and in the parasitic liver fluke *Schistosoma mansoni* 11 % are SL trans-spliced with a few genes in operons (Blumenthal, Gleason, 2003; Satou et al., 2008; Matsumoto et al., 2010; Protasio et al., 2012). Among free-living flatworms, trans-splicing was studied before (Zayas et al., 2005; Rossi et al., 2014), but there is no firm estimation of its abundance and prevalence of genes in operons. The size of operons in *M. lignano* also varies similarly to *C. elegans*, where it ranges from two to eight genes, with the most frequent intergenic distance around 100 bp, and the majority of operons comprised of two genes (see Fig. 2) (Allen et al., 2011).

The most striking finding of our study is that most of *M. lignano* SL trans-spliced genes are evolutionary conserved (see Fig. 3) and, most importantly, that overwhelming majority of neoblast-specific genes (85 %) are SL trans-spliced (see the Table). Interestingly, 39 % of neoblast genes are also clustered in operons (see the Table), suggesting their early evolutionary origin and the necessity for synchronized expression and similar transcriptional regulation. Neoblasts are the key players of outstanding regeneration capacity in free-living flatworms, and thus they are the primary subject of the studies on flatworm regeneration. All the tissue renewal and growth in adult flatworms is due to neoblast proliferation and differentiation (Egger et al., 2006; Ladurner et al., 2008; Wagner et al., 2011). Our data clearly indicates importance of SL trans-splicing for the gene regulation of neoblasts in *M. lignano* and lay ground for further studies of how exactly SL trans-splicing machinery contributes to different stages of neoblast activity.

Conclusion

Spliced leader trans-splicing affects a substantial fraction of *M. lignano* genes. We annotated and characterized the sequence of SL RNA, identified the conserved motifs at the exon-intron/exon-outron junction sites in cis- and SL trans-spliced genes, and provided the first comprehensive analysis of genes comprising operons in *M. lignano*. Most importantly, we found that the SL trans-spliced fraction is over-represented by evolutionary conserved protein-coding genes, in contrast to the non-trans-spliced part of the genome, and that the stem cell-specific genes are predominantly SL trans-spliced. Our findings suggest an important and evolutionary conserved role of SL trans-splicing in regulation and maintenance of neoblasts in *M. lignano*. Thus, a thorough investigation of the molecular mechanism of SL trans-splicing is required to fully understand the regulation of regeneration and stem cell differentiation in flatworms.

References

Allen M.A., Hillier L.W., Waterston R.H., Blumenthal T. A global analysis of *C. elegans* trans-splicing. *Genome Res.* 2011;21(2):255-264. DOI 10.1101/gr.113811.110.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.* 1990;215(3):403-410. DOI 10.1016/S0022-2836(05)80360-2.

Bailey T.L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. 2011;27(12):1653-1659. DOI 10.1093/bioinformatics/btr261.

Blumenthal T., Gleason K.S. *Caenorhabditis elegans* operons: form and function. *Nat. Rev. Genet.* 2003;4(2):110-118. DOI 10.1038/nrg995.

Boothroyd J.C., Cross G.A. Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5 end. *Gene*. 1982;20(2):281-289. DOI 10.1016/0378-1119(82)90046-4.

Conrad R., Liou R.F., Blumenthal T. Conversion of a trans-spliced *C. elegans* gene into a conventional gene by introduction of a splice donor site. *EMBO J.* 1993;12(3):1249-1255.

Danks G.B., Raasholm M., Campsteijn C., Long A.M., Manak J.R., Lenhard B., Thompson E.M. Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol. Biol. Evol.* 2015;32(3):585-599. DOI 10.1093/molbev/msu336.

Danks G., Thompson E.M. Trans-splicing in metazoans: A link to translational control? *Worm*. 2015;4(3):e1046030. DOI 10.1080/21624054.2015. 1046030.

Egger B., Ladurner P., Nimeth K., Gschwentner R., Rieger R. The regeneration capacity of the flatworm *Macrostomum lignano* – on repeated regeneration, rejuvenation, and the minimal size needed for regeneration. *Dev. Genes Evol.* 2006;216(10):565-577. DOI 10.1007/s00427-006-0069-4.

Ershov N.I., Mordvinov V.A., Prokhortchouk E.B., Pakharukova M.Y., Gunbin K.V., Ustyantsev K., Genaev M.A., Blinov A.G., Mazur A., Boulygina E., Tsygankova S., Khrameeva E., Chekanov N., Fan G., Xiao A., Zhang H., Xu X., Yang H., Solovyev V., Lee S.M.-Y.,

К.В. Устьянцев
Е.В. Березиков

Стволовые клетки и транс-сплайсинг
у регенерирующего плоского червя *Macrostomum lignano*

2021
25•1

Liu X., Afonnikov D.A., Skryabin K.G. New insights from *Opisthorchis felineus* genome: update on genomics of the epidemiologically important liver flukes. *BMC Genomics*. 2019;20(1):399. DOI 10.1186/s12864-019-5752-8.

Ganot P., Kallesøe T., Reinhardt R., Chourrout D., Thompson E.M. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol. Cell. Biol*. 2004;24(17):7795-7805. DOI 10.1128/MCB.24.17.7795-7805.2004.

Gregory T.R., Nicol J.A., Tamm H., Kullman B., Kullman K., Leitch I.J., Murray B.G., Kapraun D.F., Greilhuber J., Bennett M.D. Eukaryotic genome size databases. *Nucleic Acids Res*. 2007;35(Suppl. 1): D332-D338. DOI 10.1093/nar/gkl828.

Grudniewska M., Mouton S., Grelling M., Wolters A.H.G., Kuipers J., Giepmans B.N.G., Berezikov E. A novel flatworm-specific gene implicated in reproduction in *Macrostomum lignano*. *Sci. Rep*. 2018; 8(1):1-10. DOI 10.1038/s41598-018-21107-4.

Grudniewska M., Mouton S., Simanov D., Beltman F., Grelling M., de Mulder K., Arindrarto W., Weissert P.M., van der Elst S., Berezikov E. Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. *eLife*. 2016;5:e20607. DOI 10.7554/eLife.20607.

Guan D., McCarthy S.A., Wood J., Howe K., Wang Y., Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36(9):2896-2898. DOI 10.1093/bioinformatics/btaa025.

Hastings K.E.M. SL trans-splicing: easy come or easy go? *Trends Genet*. 2005;21(4):240-247. DOI 10.1016/j.tig.2005.02.005.

Krause M., Hirsh D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell*. 1987;49(6):753-761. DOI 10.1016/0092-8674(87)90613-1.

Ladurner P., Egger B., De Mulder K., Pfister D., Kuales G., Salvenmoser W., Schärer L. The stem cell system of the basal flatworm *Macrostomum lignano*. In: Bosch T.C.G. (Ed.). Stem Cells: From Hydra to Man. Dordrecht: Springer, Netherlands, 2008;75-94. DOI 10.1007/978-1-4020-8274-0_5.

Lasda E.L., Blumenthal T. Trans-splicing. *Wiley Interdisc. Rev. RNA*. 2011;2(3):417-434. DOI 10.1002/wrna.71.

Lei Q., Li C., Zuo Z., Huang C., Cheng H., Zhou R. Evolutionary insights into RNA trans-splicing in vertebrates. *Genome Biol. Evol*. 2016;8(3):562-577. DOI 10.1093/gbe/evw025.

Liou R.F., Blumenthal T. trans-spliced *Caenorhabditis elegans* mRNAs retain trimethylguanosine caps. *Mol. Cell. Biol*. 1990;10(4):1764-1768.

Matsumoto J., Dewar K., Wasserscheid J., Wiley G.B., Macmil S.L., Roe B.A., Zeller R.W., Satou Y., Hastings K.E.M. High-throughput sequence analysis of *Ciona intestinalis* SL trans-spliced mRNAs: Alternative expression modes and gene function correlates. *Genome Res*. 2010;20(5):636-645. DOI 10.1101/gr.100271.109.

Mouton S., Grudniewska M., Glazenburg L., Guryev V., Berezikov E. Resilience to aging in the regeneration-capable flatworm *Macrostomum lignano*. *Aging Cell*. 2018;17(3):e12739. DOI 10.1111/acel.12739.

Protasio A.V., Tsai I.J., Babbage A., Nichol S., Hunt M., Aslett M.A., Silva N.D., Velarde G.S., Anderson T.J.C., Clark R.C., Davidson C., Dillon G.P., Holroyd N.E., LoVerde P.T., Lloyd C., McQuillan J., Oliveira G., Otto T.D., Parker-Manuel S.J., Quail M.A., Wilson R.A., Zerlotini A., Dunne D.W., Berriman M. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl. Trop. Dis*. 2012;6(1):e1455. DOI 10.1371/journal.pntd.0001455.

Rossi A., Ross E.J., Jack A., Sánchez Alvarado A. Molecular cloning and characterization of SL3: A stem cell-specific SL RNA from the planarian *Schmidtea mediterranea*. *Gene*. 2014;533(1):156-167. DOI 10.1016/j.gene.2013.09.101.

Satou Y., Mineta K., Ogasawara M., Sasakura Y., Shoguchi E., Ueno K., Yamada L., Matsumoto J., Wasserscheid J., Dewar K., Wiley G.B., Macmil S.L., Roe B.A., Zeller R.W., Hastings K.E.M., Lemaire P., Lindquist E., Endo T., Hotta K., Inaba K. Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations. *Genome Biol*. 2008;9(10):R152. DOI 10.1186/gb-2008-9-10-r152.

Stover N.A., Kaye M.S., Cavalcanti A.R.O. Spliced leader trans-splicing. *Curr. Biol*. 2006;16(1):R8-R9. DOI 10.1016/j.cub.2005.12.019.

Wagner D.E., Wang I.E., Reddien P.W. Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*. 2011;332(6031):811-816. DOI 10.1126/science.1203983.

Wasik K., Gurtowski J., Zhou X., Ramos O.M., Delás M.J., Battistoni G., Demerdash O.E., Falciatori I., Vizoso D.B., Smith A.D., Ladurner P., Schärer L., McCombie W.R., Hannon G.J., Schatz M. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. *Proc. Natl. Acad. Sci. USA*. 2015;112(40): 12462-12467. DOI 10.1073/pnas.1516718112.

Wudarski J., Egger B., Ramm S.A., Schärer L., Ladurner P., Zadesenets K.S., Rubtsov N.B., Mouton S., Berezikov E. The free-living flatworm *Macrostomum lignano*. *EvoDevo*. 2020;11(1):5. DOI 10.1186/ s13227-020-00150-1.

Wudarski J., Simanov D., Ustyantsev K., de Mulder K., Grelling M., Grudniewska M., Beltman F., Glazenburg L., Demircan T., Wunderer J., Qi W., Vizoso D.B., Weissert P.M., Olivieri D., Mouton S., Guryev V., Aboobaker A., Schärer L., Ladurner P., Berezikov E. Efficient transgenesis and annotated genome sequence of the regenerative flatworm model *Macrostomum lignano*. *Nat. Commun*. 2017; 8(1):2120. DOI 10.1038/s41467-017-02214-8.

Wudarski J., Ustyantsev K., Glazenburg L., Berezikov E. Influence of temperature on development, reproduction and regeneration in the flatworm model organism, *Macrostomum lignano*. *Zool. Lett*. 2019; 5(1):7. DOI 10.1186/s40851-019-0122-6.

Xie H., Hirsh D. *In vivo* function of mutated spliced leader RNAs in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA*. 1998;95(8): 4235-4240.

Zadesenets K.S., Schärer L., Rubtsov N.B. New insights into the karyotype evolution of the free-living flatworm *Macrostomum lignano* (Platyhelminthes, Turbellaria). *Sci. Rep*. 2017;7(1):6066. DOI 10.1038/s41598-017-06498-0.

Zayas R.M., Bold T.D., Newmark P.A. Spliced-leader trans-splicing in freshwater planarians. *Mol. Biol. Evol*. 2005;22(10):2048-2054. DOI 10.1093/molbev/msi200.

Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 2003;31(13):3406-3415. DOI 10.1093/nar/gkg595.

ORCID ID

K.V. Ustyantsev orcid.org/0000-0003-4346-3868
E.V. Berezikov orcid.org/0000-0002-1145-2884