Investigation of Caenorhabditis elegans transcriptome using nanopore-based sequencing technology

Florian Bernard

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX ET DE L'UNIVERSITÉ DE TEL AVIV

ÉCOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTÉ UBX
ET ÉCOLE DOCTORALE DE NEUROSCIENCES TAU
SPÉCIALITÉ GÉNÉTIQUE

Par Florian BERNARD

INVESTIGATION OF *CAENORHABDITIS ELEGANS* TRANSCRIPTOME USING NANOPORE-BASED SEQUENCING TECHNOLOGY

Sous la direction de:
Dr. Denis DUPUY et Pr. Oded RECHAVI

Soutenue le 16 Décembre 2020

Membres du jury:

Pr. François DOIGNON	Professeur, Université de Bordeaux	Président
Pr. Peter MEISTER	Professeur, University of Bern	Rapporteur
Dr. Julian CERÓN MADRIGAL	Principal Investigator, IDIBELL	Rapporteur
Pr. Oded RECHAVI	Professeur, Tel Aviv University	Codirecteur

"The purpose of a storyteller is not to tell you how to think,
but to give you questions to think upon"

Brandon Sanderson, The way of kings

AKNOWLEDGEMENTS

First of all, I would like to thank the members of the jury, **Pr. François Doignon**, **Pr. Peter Meister** and **Dr. Julian Cerón Madrigal**, for taking the time to read and examine my work.

Then, I would like to thank all the people who have contributed, one way or another, to make these last four years an incredible adventure. I cannot list everyone but whether it was in Bordeaux or in Tel Aviv, thank you.

I particularly want to thank **Denis**, without whom none of this would have been possible. Working with you has been an amazing a very fulfilling experience, and my words alone will not suffice to say how grateful I am you chose me as your student. Thank you for your patience, for your constant support and - most of all - for always believing in me these last five years. You have helped me to grow has a researcher, but also as a person. "Every Morty needs a Rick".

The next person I want to thank is **Oded**. Thank you for giving me the incredible opportunity to work with you. You are an inspiring person, and you gather around you more inspiring persons ! If I had an excellent time in Tel Aviv and in the lab, this is mostly thanks to you.

Finally, thank you to all the person whom I had the opportunity to work with. You are now all spread around the world and it would take too much paper to thank you all personnally, but each of you participated in making the day-to-day challenges easier.

Merci à **Delphine**, **Sabrina**, **Myriam**, ainsi qu'à tous les étudiants que j'ai pu voir au labo, et aux différentes équipes de l'IECB. Un remerciement tout particulier aux membres des équipes **Mackereth**, **Innis**, **Teichmann** et **Fronze** de l'IECB.

Thank you to the entire **Rechavi lab**, past and present members. You are all incredible researchers and I am happy I can call you my friends. I particularly want to thank **Sarit** for making my arrival in Tel Aviv smoothier, even though we had to fight both french and israeli bureaucracy.
Thank you **Olga,** for the numerous coffee-breaks we had together, and **Dana** for the epic Eurovision nights ! Also, thanks to **Arielle** (chocolatine!), and **Itai** (Fourrier!) for the fun time in the lab !

Evidemment, merci à tous mes amis de l'IECB, qui ont toujours été d'un soutien inconditionnel.
Je pense notamment à **Camila**, qui a joué le rôle de la « maman scientifique » pendant de nombreuses années.
Merci à **Jean**, evidemment, qui est responsable d'avoir développé chez moi des gouts musicaux plus que douteux ! Nous n'avons pas toujours été très productifs au labo, mais c'était des moments importants pour notre santé mentale. Comme dirais Perceval et Karadoc, « Sir, on en a gros ! ».

Enfin, merci à **Alba**, **Elodie**, **Thomas** et **Guénaël**. Il est impossible pour moi de résumer tous les moments que l'on a passé ensemble en quelques lignes mais sachez que votre amitié et votre soutient m'est précieux. Si j'ai pu traverser la thèse sans encombre, c'est aussi parce que je vous savais présents à mes côtés.

Merci aussi à tous mes amis, que ce soit mon groupe d'amis d'enfances, mes amis rencontrés sur les bancs de la fac ou ceux aux terrasses des pubs Bordelais. Un merci tout particulier à **Rémy** et **Chekib**, sans qui la vie bordelaise aurait été plus morne !

Evidemment, rien de tout cela n'aurait été possible sans le soutien et l'amour de ma **famille**. Merci à mes **parents** et à ma **sœur**, qui m'ont permis d'arriver là où j'en suis aujourd'hui et qui m'ont toujours fait une confiance aveugle, même lorsque que j'ai dû prendre des décisions difficiles, pour eux comme pour moi.

Merci à mes **grands-parents maternels**, j'espère avoir pu vous rendre fier de moi.
Et une pensée aussi envers mes **grands-parents paternels**, partis trop tôt, et que je n'aurais malheureusement pas eu l'occasion de connaître.

Et puis comment ne pas remercier le reste de ma famille, notamment « **les Kikis** », et en particulier **Bastien**. Tu as longtemps été un modèle pour moi et tu continues encore aujourd'hui de m'influencer, même lorsque que plusieurs milliers de kilomètres peuvent nous séparer.

A ceux que je n'ai pas pu mentionner : **merci**.

ABSTRACT

A recent meta-analysis of alternative exon usage in *C. elegans* refined our comprehension of its transcriptome, especially regarding the splicing quantitative aspects of alternative splicing in messenger RNAs. However, Next-Generation Sequencing technologies like Illumina technology are proving to be limited to fully characterize one's transcriptome. PCR-based sequencing methods are known to introduce amplification bias affecting the overall distribution of mRNAs detected in one experiment and short-reads are not suited to accurately predict the frequency of isoforms derived from multiple alternative splicing events.

In this study, we exploited new possibilities offered by Oxford Nanopore Technology (ONT) to overcome those limitations. Nanopore-based sequencing allow us to directly sequence nucleic acids without any prior amplification step and generates long-reads covering up to the full-length of the molecule. Hence, permitting to further characterize *C. elegans* transcriptome by providing a more accurate measure of isoforms ratios, a better comprehension of exons associations during alternative splicing and by characterizing differentially trans-spliced mRNAs.

We assessed the efficiency of different sequencing kits commercialized by ONT and our results indicates that direct-cDNA sequencing is better suited for performing transcriptome analysis in *C. elegans*, in regard to the quantity and quality of data generated. Following this analysis, several direct-cDNA sequencing experiments have been performed on different populations of mRNAs: libraries of poly(A) RNAs representing the whole-animal transcriptome and libraries of SL1-enriched mRNAs. Our findings indicate that trans-spliced RNAs have an atypical behaviour during ONT's library preparation and trans-splicing of *C. elegans* mRNAs is more prevalent than previously reported. Finally, we also show that alternative promoters can lead to population of isoformes exhibiting different trans-splicing status.

Keywords: Caenorhabditis elegans ; Oxford Nanopore Technology ; RNA sequencing ; alternative splicing ; trans-splicing.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENT: **13**

Chapter 1

Introduction

RESEARCH AIMS

A recent meta-analysis of alternative exon usage in *C. elegans*, that was performed in the group of Dr. Dupuy, refined our comprehension of its transcriptome, especially regarding the quantitative aspects of alternative splicing in messenger RNAs (Tourasse et al. 2017). However, current sequencing technologies, like Illumina technology, are proving to be limited to fully characterize one's transcriptome. PCR-based sequencing methods are known to introduce amplification bias (Kebschull and Zador 2015)affecting the overall distribution of mRNAs detected in one experiment and short-reads are not suited to accurately determine the frequency of complex isoforms derived from multiple alternative splicing events.

In this study, we exploited the new possibilities offered by Oxford Nanopore Technology (ONT) to overcome such limitations. Nanopore-based sequencing allows to direct sequencing nucleic acids without any prior amplification step and generates long-reads frequently covering the full-length of the molecule. Hence, permitting to further characterize *C. elegans* transcriptome regarding isoforms ratios, exons associations during alternative splicing events and trans-splicing events.

During my thesis, I sequenced several libraries of *C. elegans* RNAs using the ONT MinION sequencer. The portability and relative ease of use of this sequencing technology made it possible for us to directly perform the acquisition of data on site. As such, and since I was responsible for designing and carrying out the different sequencing experiments, I also decided to conduct the exploration of the different datasets through bioinformatics analysis. This personally gave me the opportunity to develop my skills in both wet and dry biology but also allowed us to have a more complete and cohesive approach in this study by fine-tuning our experimental approaches according to the possibilities offered by computational biology and conversely customizing the analysis pipeline to our experimental set up and our biological questions.

At the start of this project, nanopore sequencing was not a very widespread technology, therefore we did not have access to a lot of well-established tools and methods. I, therefore, had to develop my own pipeline for the processing of the raw data into exploitable sequences. This was also the occasion for me to learn the python coding language in order to develop a number of scripts specifically adapted to answer questions of interest, such as the trans-splicing status of different *C. elegans* genes.

In the first part of this introduction is given an overview of how gene expression is regulated in eukaryotes organisms such as *Caenorhabditis elegans*. It is then followed by a closer look at two specific mechanisms working through the splicing of the messenger RNA during its maturation and responsible for the generation of different populations of transcripts: alternative-splicing and trans-splicing. In a third part, I will detail what are the advantages of using *C. elegans* as model organism for the study of complex processes such as gene expression. A review of the recent advances made towards the improvement of our comprehension of the nematode transcriptome will be addressed, along with the questions that still remains and how we can tackle current problems using new sequencing technologies. Finally, I will describe the evolution of the different sequencing technologies over the last 40 years, including the drawbacks and advantages of some of the major approaches, in order to explain why we consider Nanopore-based sequencing technology as an interesting tool for carrying transcriptome-wide studies in the worm.

1.1 - From genome to gene expression

As diverse as organisms can be, every living thing on earth possesses genetic information which reveals a common ancestry between them all. The entirety of this information constitutes what we call the genome.

The genetic material is made of deoxyribonucleic acids (DNA), composed of only four different nucleotides - adenine (A), thymine (T), guanine (G) and cytosine (C) - and the order in which those are arranged determines all the instructions necessary for the correct development, function and reproduction of every individual. Specific regions in the genome encoding for those instructions are called genes, and the process by which this information is used by the cell to produce proteins is referred as gene expression.

1.1.1 - Overview of genome organization and maintenance

a) Genome organization

In eukaryotic cells, the genetic material is stored inside the nucleus. The role of this structure is to isolate and protect the genetic material from the content of the cytoplasm thanks to a nuclear envelope made of two lipidic bilayers. The nuclear envelope also presents nuclear pores which allow communication with the other compartments of the cell. Inside the nucleus, the DNA is wrapped around histone proteins which helps to ensure the compaction of the genetic material into a structure called the chromatin (Holde 1988).

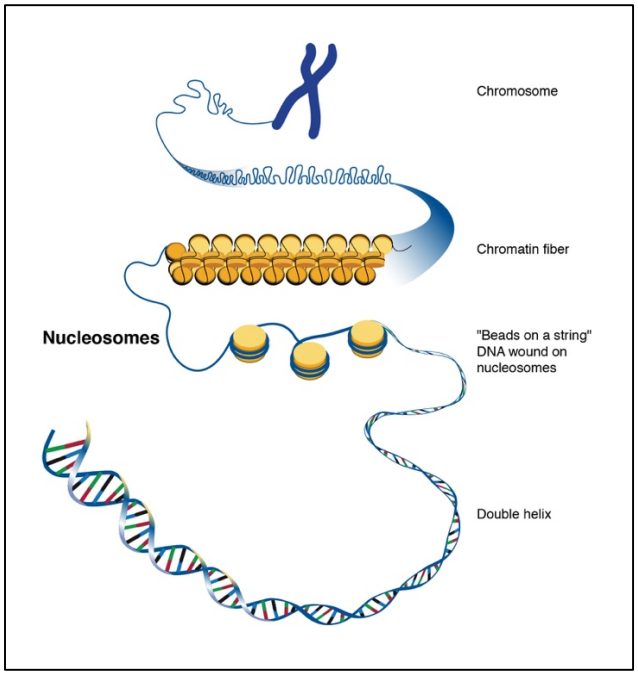


Figure 1 - Organization of DNA inside the nucleus (From the National Genome Institute).

This structure is highly dynamic and allow to regulate the accessibility of DNA to proteins depending of cellular status. When needed, like during cellular division, the chromatin can be further condensed into chromosomes. Altogether, the structure of the nucleus and its inner organization permits to finely controls cellular activity, through the regulation of gene expression.

b) Genome maintenance

During replication, a helicase protein is responsible for unwinding DNA, locally separating the two single strands of DNA. This allow a large enzymatic complex (called DNA polymerase) to bind to each strand to perform DNA synthesis. During this process, the DNA polymerase use the original strand of DNA as a template to synthesize a completely new strand. As the helicase unwinds both strands, this phenomenon happens simultaneously on both strand, resulting in two double-stranded DNA molecules, each one made up of one of the initial strand as well as a newly synthesized strand.

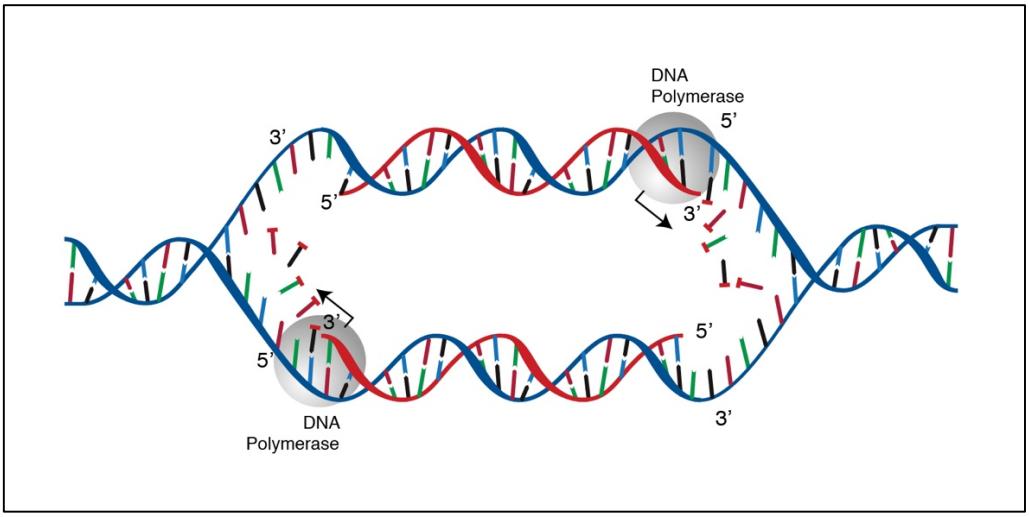


Figure 2 - Schematic of DNA replication (National Genome Institute).

With the development of techniques to purify specific enzymes, this naturally-occurring process have been extensively used by researchers. First, in Polymerase Chain Reactions (PCR) to amplify molecules of DNA in vitro, and later in different sequencing technology for deciphering nucleic acid sequences.

c) genome content

A genome's content is usually classified into coding and non-coding regions. Coding regions are carrying sequences that will be used to produced proteins while non-coding regions used to be referred as "junk" or "selfish" DNA (Doolittle and Sapienza 1980; Orgel and Crick 1980)
However, it has now been shown that non-coding RNA plays a critical role in several regulatory functions (Werren 2011).

With the development of the first sequencing technologies in the 1980's, researchers have been able to sequence the genome of several species. In 1999, *Caenorhabditis elegans* was the first metazoan whose genome was fully sequenced (Consortium* 1998). This project permitted researchers to determine that *C. elegans* possess ~20 000 genes distributed across six pairs of chromosomes for a total size of about 100Mb. Surprisingly, upon completion of the Human Genome Project (HGP) in 2003, it was found that the human genome also contained ~20 000 genes for a total genome size of 3.2Gb(Consortium 2001). This discovery lead researchers to further investigate how a small organism like *C. elegans* could have the same number of genes as a much more complex organism like humans. Since then, several advances have been and we now know that gene expression is a heavily regulated process and that different mechanisms are able to modulate the function of a gene.

1.1.2 - Gene expression

Because it is important to protect the information stored inside the nucleus, and because not all of it is needed all the time, an intermediary molecule is responsible for carrying the message from the nucleus to the cytoplasm. During this process, the information is passed onto a molecule of ribonucleic acid (RNA) that will be used as template by the cellular machinery to produce proteins. This mechanism, called RNA transcription, is the primordial step of what we call gene expression.

By regulating gene expression, a cell controls the timing and the amount of production of any functional gene product, which gives the cells the ability to react and to adapt to various stimuli. Furthermore, gene regulation is at the basis of development by fine-tuning the expression of all the genes required for the various differentiation events necessary to produce a viable organism.

a) Mechanism of RNA transcription

Different RNA polymerases (RNAP) - complexes made up of several subunits - are responsible for the transcription of different class of RNAs:

- RNA polymerase I: transcription of most ribosomal RNAs (rRNAs), with the exception of 5S (see below) (Drygin et al. 2010).
- RNA polymerase II: transcription of messenger RNAs (mRNAs) and most of the small nuclear RNAs (snRNAs) (TAMURA et al. 1996)
- RNA polymerase III: transcription of transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), micro RNAs, as well as U6 snRNA, 7SK snRNA and 5S rRNA (Dieci et al. 2007)
- RNA polymerase IV and V: transcription of small interfering RNAs (only in plants) (Herr et al. 2005).

The process of RNA transcription is divided into three major steps (Buratowski et al. 1989; Gnatt et al. 2001):

1. Initiation:

The first step in RNA transcription involves the interaction of the RNA polymerase with several transcription factors, called the pre-initiation complex. This complex allows binding onto the DNA strand in a region that is termed promoter region.
At the end of the initiation, the polymerase complex is still prone to premature arrest. Many abortive cycles can happen before the RNA polymerase is finally able to synthesize an RNA fragment long enough (10-14bases) to reach its exit channel, which in turns triggers promoter escape and the start of the elongation (Dvir et al. 2001).

2. Elongation:

During RNA synthesis, one strand of the DNA is used as a template. It is referred as the non-coding strand. The RNA polymerase uses base complementarity to insert the correct ribonucleotide into the newly synthesized RNA, resulting in a perfect copy of the coding strand – save for the insertion of uracyl (U) instead of thymine (T).

3. Termination:

This termination step is different for all three eukaryotic RNA polymerases and not very well understood for RNA polymerase III. In protein-coding genes, the end of the transcription is coupled with the polyadenylation of the mRNA (Minvielle-Sebastia and Keller 1999; Yonaha and Proudfoot 2000). As the RNA pol II passes a specific signal, the nascent RNA is cleaved from the RNA polymerase to be fully matured. The remaining strand, still attached to the polymerase, is then digested by a 5' exonuclease. Once the 5' exonuclease reaches the polymerase, it triggers the end of the transcription and the RNA polymerase gets released from the DNA.

b) Maturation of messenger RNAs

The maturation of the molecule of mRNA is a critical step that happens co-transcriptionally. It involves the addition of a cap on the 5' extremity (Banerjee 1980) and the addition of a poly-(A) tail on the 3' extremity (Bachvarova 1992; Lewis et al. 1995). Additionally, the splicing of the nascent RNA allows to remove non-coding regions from the sequence, while adjoining the coding regions necessary to produce the protein.

This set of modifications is also critical for the export of the RNA to the cytoplasm, its protection from ribonucleases and the recruitment of other protein complexes that are essential to the proceeding of downstream processes (Gallie 1991).

1. RNA capping

In most organisms, RNA capping lead to the addition of a 7-methylguanosine cap on the 5' end of the mRNA. However, in *C. elegans*, the addition of the cap on the 5' end of the mRNAs is performed through an additional step that is called trans-splicing (see **section 1.2**). This mechanism is responsible for the replacement of the 5' UTR region by a 22nt sequence that carries a 5' trimethylguanosine (TMG) cap (Thomas et al. 1988; Liou and Blumenthal 1990).

The addition of the cap to the 5' extremity allow the recruitment of cap-binding proteins (CBP) that help to prevent the degradation of the mRNA and allow its export to the nucleus (Daneholt 1997). After going through the first round of translation, CBP are replaced by translation factors.

2. RNA splicing

In higher organisms, genes are discontinuous, they are made of coding regions (exons), separated by non-coding regions (introns). This discovery was possible thanks to the work of Richard J. Roberts and Phillip A. Sharp, who both demonstrated that the RNA molecule corresponds to several well-defined segments of the DNA molecule it originates from (Berget et al. 1977; Chow et al. 1977), rather than being an exact copy. This also led to the realization that RNA molecule must be processed by the cell before being fully functional. The process of removing all the introns from the sequence while ligating the exons together is called RNA splicing.

The splicing of the molecule happens in the nucleus in parallel of the transcription and finishes once the RNA is fully transcribed (Singh and Padgett 2009). The different reactions leading to intron removal are catalyzed by the components of the spliceosome, a complex which involves five small nuclear ribonucleoproteins (snRNPs): U1, U2, U4/U6 and U5(Bringmann et al. 1983). snRNPs are made of the association of a small nuclear RNA - about 150bp long and rich in uracyl - with RNA-binding proteins.

The splicing of an intron works as follow:

1) The U1 snRNP binds to the 5' extremity of the intron (GU).
2) The U2 snRNP binds to the branching point, an adenine (A) situated 20 to 40bp downstream of the 3' extremity of the intron (AG) (Burge et al. 1999). This interaction is facilitated by the recruitment of U2AF factor at the 3' splice site (AG).
3) The RNA molecule is flexed near the branching site, which puts U1 and U2 snRNPs near each other and allow recruitment of U4/U6 and U5. This complex is able to catalyze the first trans-esterification reaction between the 5' splice site and the branch-site (Moore and Sharp 1993).
4) U4 and U1 snRNPs are then released from the complex and some internal rearrangements happens, allowing the U2/U5/U6 complex to catalyze another trans-esterification reaction between the 3' splice site and the OH residue of the 5' splice-site.
5) Following the second reaction, both exons end up ligated together and the intron is released as a lariat (Padgett et al. 1984), along with the remaining snRNPs that will be recycled by the cellular machinery.

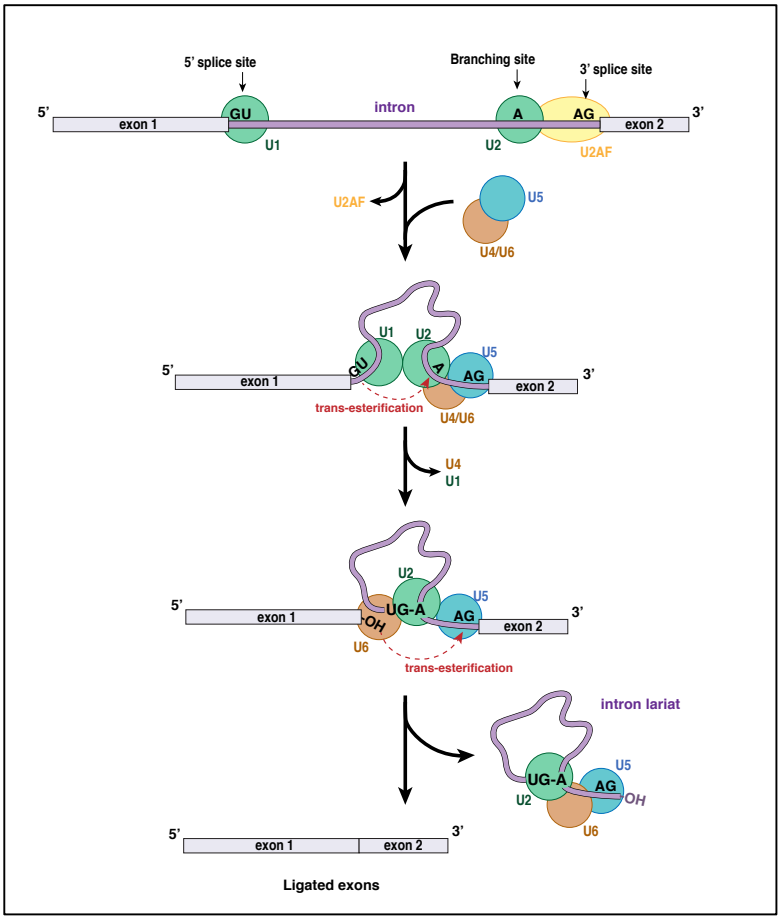


Figure 3 - Description of the different steps leading to the removal of an intronic sequence within a messenger RNA molecule.

Once an intron is removed from the sequence, a protein complex named the exon-junction complex (ECJ) is deposited 20 to 24 nucleotides upstream the exon-junction (Hir et al. 2001). Its binding to the RNA is made independently from the sequence. The EJC plays an important role in the localization of the mRNA through its interaction with the proteins of the nuclear pore and serves as a binding platform for other factors present in the nucleus (Luo and Reed 1999; Zhou et al. 2000).

This complex plays a critical role in mRNA surveillance. During translation by the ribosome, the EJC gets removed from the molecule (Ishigaki et al. 2001). However, if a premature stop codon is found, the ribosomal complex will detach from the mRNA and downstream EJC will not get released, triggering the degradation of the mRNA via the non-mediated decay (NMD) pathway (Hir et al. 2001).

3. RNA polyadenylation

The end of the transcription is signaled by the presence of a polyadenylation sequence (AAUAAA) in the 3' untranslated region (UTR) of the mRNA (Hashimoto and Steitz 1986). When the RNA

polymerase encounters this sequence, there is recruitment of the cleavage and polyadenylation specificity factors (CPSF) complex (Barabino et al. 1997). The CPSF complex will cut the nascent RNA just downstream of the polyadenylation signal sequence and recruit a polyadenylate polymerase whose role is to add adenine monophosphate units to the 3' end. This results in a poly(A) tail of approximately 250bp on the 3' extremity of the mRNA (Bienroth et al. 1993; Wahle et al. 1993).

The poly(A) tail is an important feature of mRNAs as it allows the recruitment of various poly(A)-binding proteins (PABP) which helps the stability of the molecule and its export to the nucleus, among other functions (Colgan and Manley 1997).

1.1.3 - Different means of regulation

As mentioned previously, gene expression needs to be regulated to allow the cell to adapt to various stimuli and to fine tune its activity. Regulation can be achieved at different level by affecting DNA, RNA or the protein itself. Some of the main processes involved are described below.

a) Epigenetics modifications at the level of DNA

One of the first possibilities to affect gene expression is through chromatin remodeling.

1) DNA methylation: in eukaryotes, 5-methylcytosines are common modifications found at CpG dinucleotides. Several studies have shown that methylation of CpG in promoter regions was negatively correlated with gene expression (Deaton and Bird 2011).

2) Histone modifications: the level of condensation of the chromatin can be altered by adding reversible modifications to the histones. This includes methylation, acetylation, phosphorylation, etc. This is referred as the "histone code" (Jenuwein and Allis 2001). Those modifications affect the state of the chromatin and can lead to regions being made more accessible to transcription factors, or on the contrary to regions being made unavailable.

Long non-coding RNAs: They have been found to interact with several genomic regions, acting as a scaffold for the recruitment of specific protein complex and helping in regulating gene expression(Tsai et al. 2010)

b) Post-transcriptional regulation

A good way to regulate gene expression is by directly affecting the mRNA molecule itself. As seen with the maturation of mRNA, such modifications positively affect the stability or the localization of the molecule and are critical for its function. RNA-binding factors can also affect its activity by competing with other factors or enhancing/blocking possible interactions.

A distinctive set of post-transcriptional regulatory mechanisms, working via RNA splicing, will be further detailed in the second section of the introduction.

c) Post-translational regulation

After the protein is produced, it is still possible to regulate its activity through the addition or removal of specific molecules.
One such example is the addition of phosphorylation marks by kinase proteins (BURNETT and KENNEDY 1954) . This modification can change the conformation of numerous proteins which, in turn, leads to major functional changes. Changes in conformation can affect the activity of the protein (inhibition or increased activity) and their interactions with other partners. It can also change the localization of certain proteins, such as transcription factors. Furthermore, phosphorylation marks can be removed by phosphatase proteins, allowing this modification to act as an on/off switch that can be controlled by the cellular machinery to react to environmental changes (Hunter 1995) .
Other post-translational modifications can also affect the stability of the protein and target it for degradation (ubiquitinylation, SUMOylation, etc). Similarly to phosphorylation, they have an effect on the localization of the protein by targeting it to specific compartments of the cells or for cellular excretion. Finally, they can directly change its activity by reducing – or even preventing – its interaction with other protein complexes (Nussinov et al. 2012).

d) Codon bias

Unlike the three modes of regulation presented above, codon bias is not a mechanism that can be regulated by the cell as it is directly encoded in the genome and, therefore, does not change over time, however it can still affect the level of gene product.

During translation, the order of the amino-acids in the nascent protein chain is determined by triplets of letters - referred as codons - in the RNA sequence. The four different letters (A, U, G and C) can be in any of the three positions of a codon, which makes a total of 64 unique codons. Yet, cells only use 20 different amino-acids for protein synthesis. This is explained by the fact that several codons are used for the same amino-acid. This redundancy of the genetic code allows some mutations to be "silent" as a single nucleotide change will not always affect the resulting protein.

Nonetheless, not all codons encoding the same amino-acid are used with the same frequency and the tRNAs responsible for carrying the correct amino-acids are not present at the same level (Sharp and Li 1987). Hence, a point mutation can still affect protein synthesis rate if a less frequent codon needs to be used by the ribosomes. Moreover, different species exhibits different codon bias. This bias needs to be addressed when expressing a foreign gene - like the Green Fluorescent Protein (GFP) that was identified in jellyfish - in another organism.

1.2 - Regulating gene expression through RNA splicing

As seen in the previous section, RNA splicing is an essential process that takes place during RNA maturation. Similarly to other mechanisms, it can be heavily regulated by the nuclear and cellular machinery, but due to the inherent nature of the splicing process, regulating RNA splicing is a critical step that can affect the nature of the encoded protein.

In this section are described two different aspects of the splicing of RNA molecules in *Caenorhabditis elegans*.

1.2.1 - Alternative splicing

The first evidences of alternative splicing dates to 1977, at the same time of the discovery of RNA splicing by Roberts and Sharp. By studying adenovirus RNAs, they observed that primary RNA transcript could be spliced in different ways, allowing for the mRNA to encode different viral proteins (Berget and Sharp 1977).

However, the interest for the study of alternative splice has increased over the last 20 years with the completion of different animal genomes. The discovery that complex organism like humans could have about the same number of genes like *C. elegans* led researchers to wonder how the observed complexity of humans could results from such a limited set of genes. One answer to this paradox is the possibility for a gene to encode more than one protein. In humans, alternative splicing is now thought to affect between 50 and 75% of all protein coding genes.

a) Different types of events

Under the name "alternative splicing" is regrouped a set of post-transcriptional events that can happen during the splicing of the mRNA molecule and that ultimately affects the final sequence of the matured RNA. As it became easier to sequence nucleic acids, it has been possible to determine different categories of alternative splicing events (**Figure 4**) (Blencowe 2006; Calarco et al. 2007).

- Exon cassette: removal of an exon during splicing of the molecule.
- Mutually exclusive exons: the splicing of an exon causes the removal of another exon.
- Intron retention: the sequence of an intron is kept in the final sequence.
- Alternative 5' or 3' splicing sites: different splices sites can be preferentially used.
- Alternative promoters: different promoter regions are available.
- Alternative polyadenylation sites: different polyadenylation sites are available (Edwalds-Gilbert et al. 1997).

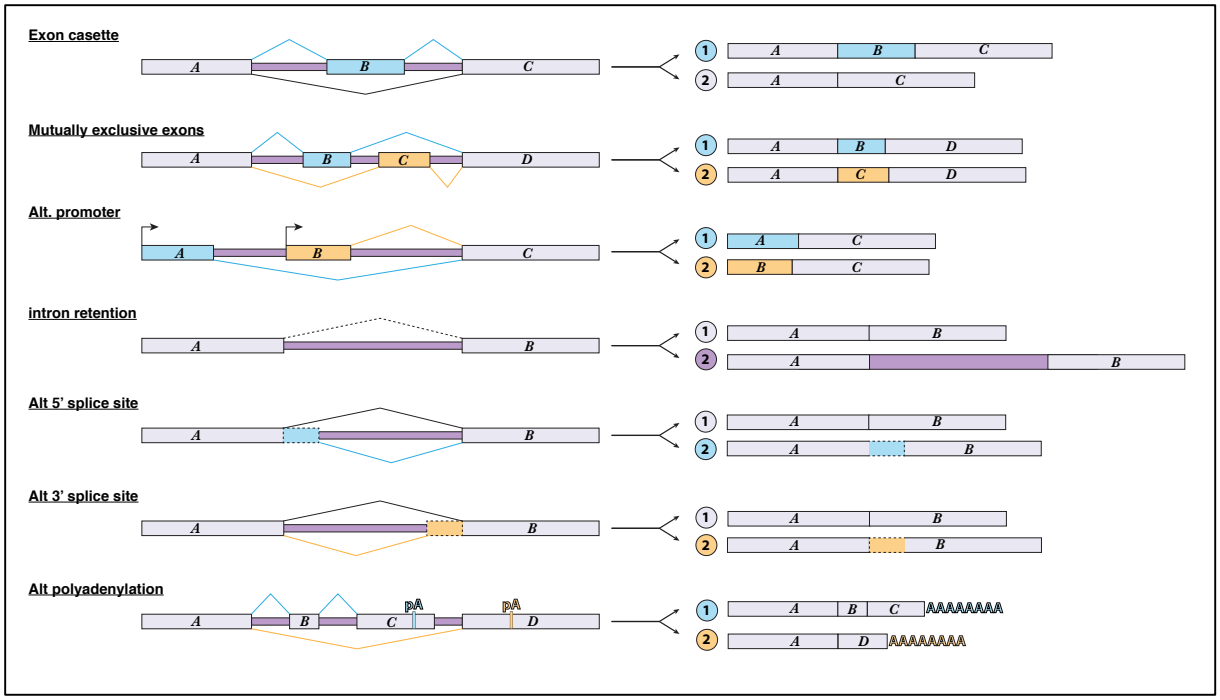


Figure 4 - Different types of alternative splicing events. The resulting isoforms for each event are depicted on the right.

b) Alternative splicing is determined by a "splicing code"

During the process of alternative splicing, the exons that ends up in the mature mRNA are entirely defined by their interaction with cis- and trans- acting factors. These interactions are the basis of what we call "the splicing code".

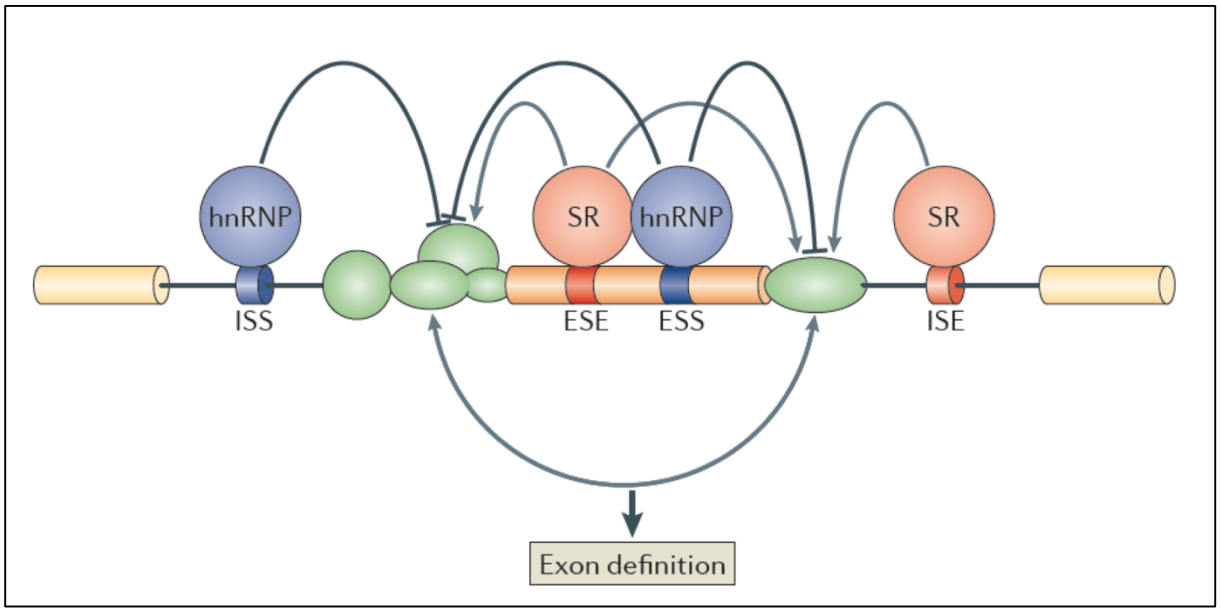


Figure 5 - the splicing code is determined by cis- and trans-acting factors.

Cis-factors:

The splicing of a pre-mRNA requires three different types of elements: a 5' splice site, a 3' splice site and a branching point. Those sites allow to determine where the core-components of the spliceosome will interact with the pre-mRNA.

The presence of specific cis-regulatory sequence on the pre-mRNAs can determine the fate of the pre-mRNA (Wang and Burge 2008). Those factors include Exonic Splicing enhancer (ESEs) and Intronic Splicing Enhancers (ISE) on which positive trans-acting factors can bind. On the other hand, Exonic Splicing Silencers (ESSs) and Intronic Splicing Silencers (ISSs) are bound by negative trans-acting factors.

Once the trans-acting factors are recruited, they can interact with the spliceosome machinery in order to affect the sequences that will remain in the matured mRNA.
In general, cis-acting elements function additively. ESEs and ISEs tend to play an important role in constitutive splicing, while ESSs and ISSs are more important for the control of alternative splicing events (Wang and Burge 2008).

Trans-factors:

The two main types of trans-acting proteins are the Serine-Arginine proteins (SR) and the Heterogeneous ribonucleoprotein particle (hnRNPs). SR proteins are majorly found associated with enhancer cis-factors, while hnRNPs are mostly found associated with silencers cis-factors.

SR proteins: In addition to the coupling of SR proteins to enhancers cis-factors, SR proteins can interact with U1 snRNP and a subunit of U2AF (Jeong 2017). SR proteins generally contain one or two RNA binding domains and a C-terminal domain containing a Serine-Arginine rich domain. SR proteins combine with SR-like proteins to select exon splicing enhancers on RNA transcripts causing U2 snRNP to bind to the upstream, adjacent branch site and causing spliceosome assembly at the specific 3' site selected by the SR proteins (Blencowe et al. 1999)

hnRNP proteins: these proteins are multifunctional and participates in all crucial aspects of RNA processing, including pre-mRNA splicing, mRNA export, localization, translation, and stability. Furthermore, they are highly conserved from nematodes to mammals.
One of the most studied hnRNP is hnRNP A1, which favour exon skipping in several mRNAs, including its own. The main function of this protein is thought to be to antagonize SR proteins. The result of this competition between hnRNPs and the SR proteins is heavily on the concentration of each proteins within the cell (Mayeda and Krainer 1992; Mayeda et al. 1994).

c) The importance of alternative-splicing

Over the years, numerous studies have shown the importance of alternative splicing in several biological processes.

One such example is the mechanism of sex determination in *Drosophila melanogaster*.
Sex-lethal (Sxl) is an RNA binding protein that is required for induction of female sexual identity in both somatic and germline cells. During embryonic development, the number of copies of chromosome X controls the sexual fate of the individual (Steinmann-Zwicky et al. 1990). The presence of two copies of chromosome X triggers the activation of gene *Sex-lethal* (Sxl). Following protein synthesis, the SXL protein can control splicing of two different target: gene *transformer* (*tra*) and itself, through an auto-regulatory feedback loop (Bell et al. 1988) .

In *tra* pre-mRNA, SXL repress the use of a non-sex-specific 3' splice site, therefore forcing the use of an alternative, female-specific, splice site. This is mediated by binding of SXL to the polypyrimidine tract associated with the non-sex-specific site, which then prevents binding of U2AF and redirects it to the alternative 3' splice site. This female specific isoform produces a functional TRA protein which controls alternative splicing of its downstream target.

Additionally, SXL induces exon skipping of a male-specific exon in its own pre-mRNA. The presence of a stop codon within the sequence of this male-specific exon gives rise to a non-functional protein. Thus, SXL-dependent exon skipping in females establish a positive feedback loop that is necessary to maintain female differentiation.

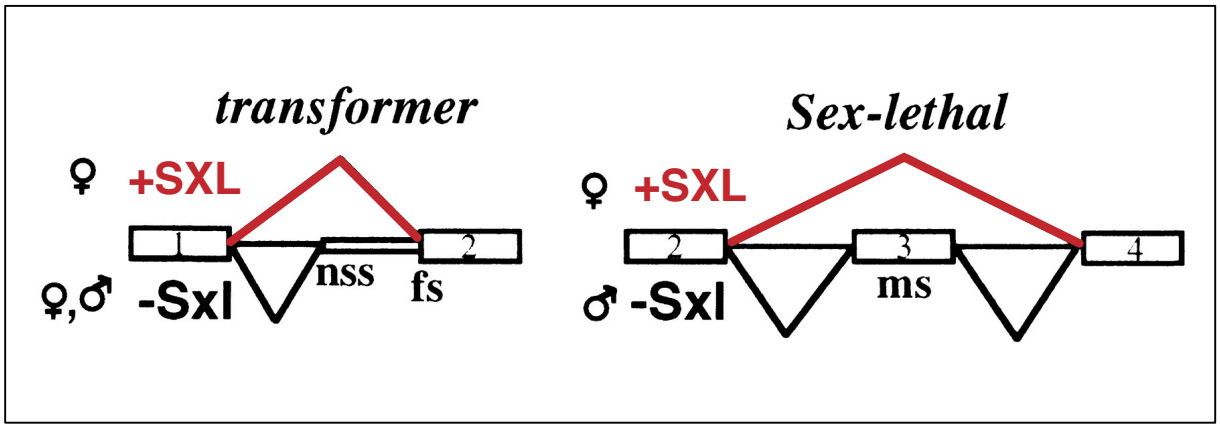


Figure 6 - SXL maintains sexual determinism by triggering sex-specific alternative splicing events. In *transformer* pre-mRNA, SXL prevents the use of a non-sex-specific (nss) 3' splice site and forces the use of a female-specific (fs) 3' splice site. In its own pre-mRNA, SXL induce skipping of exon 3, a male-specific exon (ms) containing a stop codon. The resulting isoforms lacking exon 3 produces a functional protein that can maintain sexual determinism. Female-specific alternative events are highlighted in red.

In *C. elegans*, alternative splicing also plays a role in sex determination. In hermaphrodites, *fox-1* gene (feminizing locus on chromosome X) encodes a RNA Recognition Motif (RRM) protein that can inhibit post-transcriptionally the expression of *xol-1* gene - the major specifier of male fate - by inhibiting splicing of the terminal intron of xol-1 pre-mRNA and prevents production of a functional XOL-1 protein. (Nicoll et al., 1997)

Tissue-specific expression of alternative splicing factors has also been demonstrated in *C. elegans*. SUP-12 is a splicing factor that was identified as a genetic suppressor of unc-60 mutations (Anyanful et al. 2004). unc-60 is a gene that can produce two different isoforms: unc-60a and unc-60b. Their respective proteins, UNC-60A and UNC-60B, are respectively expressed in non-muscle cells and muscle cells. It was shown the expression of sup-12 in muscle cells is required for the alternative splicing of unc-60b in this tissue. Loss of sup-12 expression leads to the production of UNC-60A in muscles.

1.2.2 - Trans-splicing

The trans-splicing is a mechanism in which a specific sequence of RNA - called splice leader (SL) sequence - is added onto the 5' extremity of multiple mRNAs.

a) Historical findings and conservation across species

The mechanism of trans-splicing was first discovered in Trypanosomes, a parasitic organism in which a SL sequence was detected during the characterization of cDNAs clones encoding surface glycoproteins (VSGs) (Sutton and Boothroyd 1986). In trypanosomes, gene expression is polycistronic and the resulting long RNAs needs to be matured into monocistronic units before being fully functional. It has now been shown that coupling of trans-splicing and polyadenylation is responsible for the maturation of the RNAs in Trypanosomes(Preußer et al. 2012) . In this species, trans-splicing is the only splicing and all the mRNAs begin with a SL sequence of 39nt derived from the 5' end of a 137nt RNA called medRNA.

After the discovery of trans-splicing in trypanosomes, it was found to exist in other metazoans, including cnidarians, ctenophores, rotifers, flatworms, nematodes, crustaceans and sponges. However, trans-splicing has not been found in any plants, fungi and insects or vertebrates (**Figure 7**).

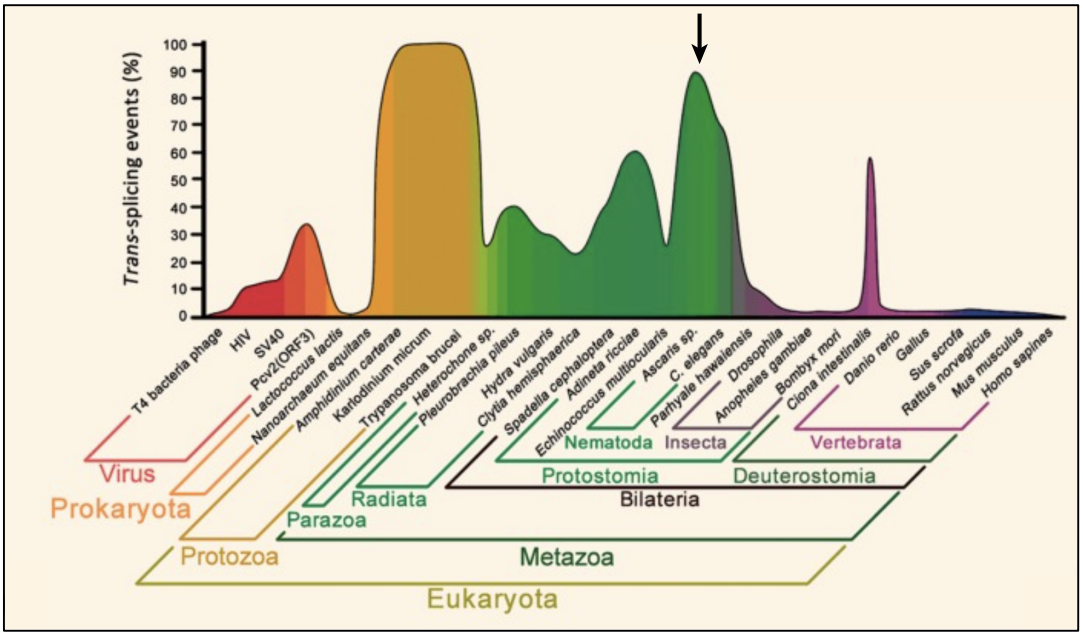


Figure 7 - Conservation of trans-splicing across the tree of life. The black arrow indicates the nematode group. **(From Lei et al, 2016)**

b) Trans-splicing mechanism in *C. elegans*

In *C. elegans*, trans-splicing was initially uncovered during the study of the 5' extremity of the actin gene where it was reported the presence of a 22nt sequence on the mRNAs (Krause and Hirsh 1987). This first SL sequence (SL1) is donated by a 100nt small nuclear ribonucleoprotein particle (snRNP) (**Figure 8.A**).

This process is closely related to cis-splicing (intron removal) where the 5' splice site is on the SL RNA, and the site of SL addition (3' splice site) is on the pre-mRNA (Kent and Zahler 2000).
The reaction happens through a branched intermediate, similar to the lariat of cis-splicing: cleavage of the 5' exon and of the SL sequence, and formation of an intermediate between the two (**Figure 8.B**). Finally, splicing of the SL to the first exon of the pre-mRNA in the second step. The 5' region of the pre-mRNA is called the outron (Conrad et al. 1991).

However, we still do not know how the SL snRNP recognizes the 3' trans-splice site of pre-mRNAs. Several hypotheses have been advanced. One of them postulates that the interaction of the U2 snRNP with the 3' splice site without matching donor site recruits the SL1 snRNP when it cannot interact with any upstream U1 snRNP, but this has not been demonstrated. Another possibility is that the SL snRNP is attracted to the 5' end of the pre-mRNA by the RNA polymerase II complex.

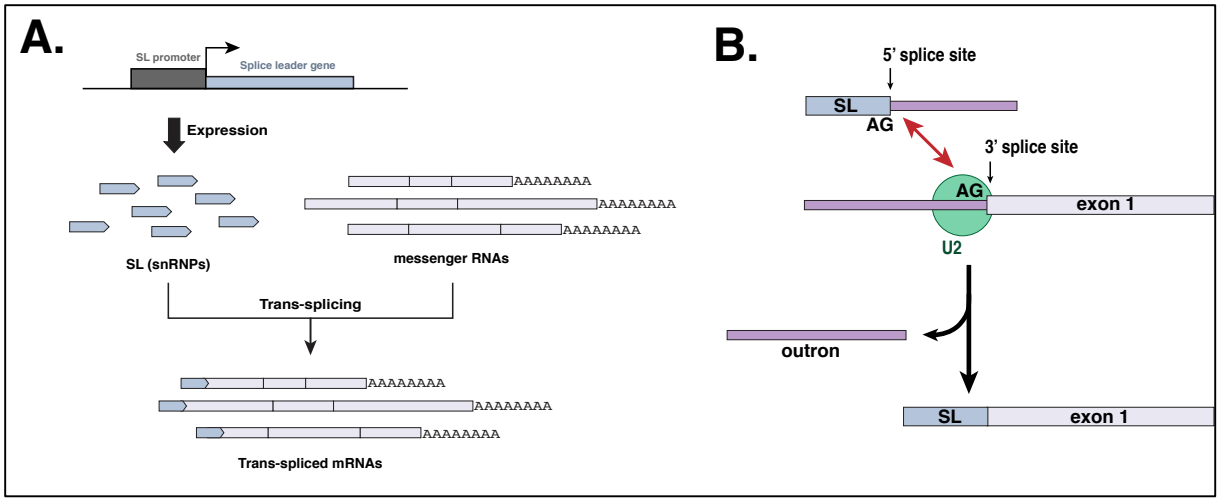


Figure 8 - Trans-splicing in *C. elegans*. A) General mechanism of trans-splicing. **B)** Interaction the SL snRNP with the 3' splice site of the pre-mRNA. Following trans-splicing, the outron is removed from the sequence of the messenger and the SL sequence is added.

c) Identification of a second SL sequence

In 1989, a second SL sequence (SL2) was found. This sequence is the same size of the first SL sequence identified (SL1), however it exhibits a different sequence. This sequence was initially found at the 5' end of the *gpd-3* gene.

The *gpd-3* gene, and the others that have been found to receive a SL2 sequence on their 5' extremity, are organized within clusters of genes with the same orientation on the genome (Spieth et al. 1993; Zorio et al. 1994). The *gpd-3* gene is the last gene of a three-gene cluster (*mai-1* / *gdp-2* / *gdp-3*) and both downstream products exhibits mRNAs trans-spliced with SL2. The first gene of this cluster (*mai-1*) is not found trans-spliced, however, many first-genes in such clusters are found associated with SL1.

Since the original discovery of this cluster, many more mRNAs have been found associated with SL2 sequences and many gene's clusters have been found. A microarray analysis of the entire genome has demonstrated how robust the correlation between gene's cluster and SL2-containing gene is (Blumenthal et al. 2002). In this study, the authors have identified more than 1 000 clusters for which downstream messengers are trans-spliced to SL2. These clusters contain more than 2 600 genes.

The remaining question is how genomic position can affect trans-splicing specificity. It was hypothesized that gene clusters acts in a similar way as bacterial operons, with the transcription of the entire cluster by a promoter sequence situated on the 5' extremity of the region. Yet, in bacteria,

mRNAs from operon regions are polycistronic, where in *C. elegans* they are processed into monocistronic units. This is because cleavage and polyadenylation occurs on the 3' end of the upstream gene, and is accompanied by the addition of a SL2 sequence on the downstream gene, in a similar fashion as to how trans-splicing and polyadenylation are coupled in trypanosomes.

This hypothesis is supported by several experimental observations (Spieth et al. 1993):

- Evidences of co-transcription have been demonstrated: cDNAs clones which contained the sequence of *mai-1* and *gdp-2* (the two other genes from *gdp-3* cluster) were isolated and the intergenic sequence between them was found.

- When expressing a construct containing the *gdp-2* and *gdp-3* gene pair under the control of a heat shock promoter, expression of the downstream gene (*gdp-3*) is dependent on heat shock and its product is trans-spliced to SL2. Furthermore, SL2 trans-splicing is dependent on the promoter being located upstream of the first gene.

- Mutation on the poly(A) site of the upstream gene leads to accumulation of polycistronic sequences, indicating that processing of polycistronic sequences is directly linked to polyadenalyation of the upstream gene product.

- The insertion of a gene normally trans-spliced by SL1 between the *gdp-2* and *gdp-3* genes, leads to its products receiving primarily SL2 which indicates that being a downstream gene inside an operon is sufficient for triggering SL2-specific trans-splicing.

d) Function of trans-splicing

Trans-splicing is a very conserved mechanism amongst nematodes and there is a striking conservation of the SL sequences found attached to the messenger RNAs, even though downstream sequences of the SL gene (region that is not trans-spliced) have diverged. Yet, the role of splice leader sequences is not known.
It is likely this sequence plays a role in translation initiation. Furthermore, SL snRNPs carry a trimethylated (TMG) cap that remains attached to the mRNAs after trans-splicing. In mammalian extracts, a TMG cap is known to inhibit translation, however, in *C. elegans*, it has been shown this specific cap stimulates translation activity when positioned on the 5' extremity of the SL sequence (Maroney et al. 1995; Lall et al. 2004). It was also shown that a variant of the cap binding translation initiation factor (eIF4E) could recognize the TMG cap (Keiper et al. 2000).

While we do not know exactly the role of the SL sequence, trans-splicing has been shown to be essential for viability (Ferguson et al. 1996). An embryonic lethal mutation in the *rrs-1* gene cluster is a deletion of all tandem copies of the 1Kb sequence that encode a 5S ribosomal RNA and the

SL1 RNA. This lethality can be rescued by the expression of an extrachromosomal array carrying the SL1 gene alone.

Nowadays, trans-splicing has been extensively found in many genes of *C. elegans* and, while early findings reported that 70% of all protein coding genes exhibited trans-splicing, recent work from our team determined that at least ~86% of *C. elegans* genes are concerned with clues supporting that the process is indeed ubiquitous (Tourasse et al. 2017). Continued discovery of new trans-spliced genes, thanks to the use of more sensitive approaches for transcriptomics analysis, along with the ubiquitous nature of the system in organism such as trypanosomes, also indicates that trans-splicing in *C. elegans* might be more prevalent than initially reported.

1.3 - *Caenorhabditis elegans* as a model organism for genetic studies

1.3.1 - Generalities

Caenorhabditis elegans is a roundworm, member of the phylum Nematoda, which lives in the soil and feeds on bacteria. It was introduced as a model organism by Sydney Brenner, in 1960, and is nowadays one of the most studied model organism in modern biology (Brenner 1974).

b) Life cycle

During its life, each hermaphrodite can produce a progeny of ~300 individuals by self-fertilization. This number can go up to a thousand individuals upon being fertilized by a male - the limiting factor becoming the number of eggs a hermaphrodite is able to produce.

The embryogenesis process takes about 16 hours at 20°C. After fertilization, the embryo is protected by an impermeable eggshell which allows it to continue development independently from the mother. However, the eggs are generally kept *in utero* until they reach the 30-cells stage. After egg laying, development continues *ex utero* until hatching.

During its development, *C. elegans* passes through four different larval stages (L1, L2, L3 and L4) before reaching full adulthood. The passage from one stage to another is characterized by a sleep-like period of inactivity called lethargus (Cassada and Russell 1975), during which a new cuticle is being produced. This state ends with the molting of the old cuticle.

The development of *C. elegans* is a temperature dependent process. At 20°C, it generally takes ~128 hours for a worm to reach the adult stage. However, development at lower temperatures (15°C) will increase this time (~90H) while higher temperature (25°C) will considerably shorten it (~47H). Nonetheless, the temperature at which development occurs has no effect on the morphology.

When confronted to adverse conditions (starvation, high temperatures, etc.), *C. elegans* L2 can enters a specific stage called the dauer stage in which the cuticle surrounds the animal (Cassada and Russell 1975). The cuticle then prevents the worms to eat and leads to developmental arrest but also provides an enhanced resistance to chemicals and environmental stresses. In this stage, dauer worms can survive for a long time. This is the stage in which *C. elegans* are most often found in the wild. Upon reintroduction of a source of food, the dauer larva loses its cuticle by molting and can then restart its development.

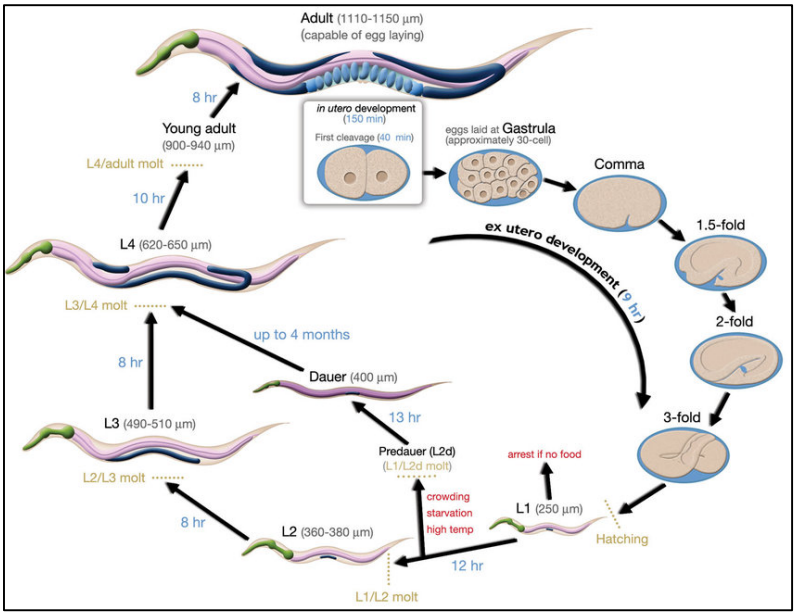


Figure 9 - *Caenorhabditis elegans* life cycle (Wormatlas.com).

c) sexual dimorphism

C. elegans worms are usually found as hermaphrodites (XX) but their populations usually contain 0.1-0.2% of males (XO), due to a rare meiotic non-disjunction of the X chromosome. Both sexes present morphological differences that makes them easily recognizable by simple observation under a binocular. Since males only produce sperm they are often thinner than their counterpart due to the small size of their gonads and the absence of eggs. Additionally, they present a very characteristic tail that is flattened into a fan (**Figure 10**).

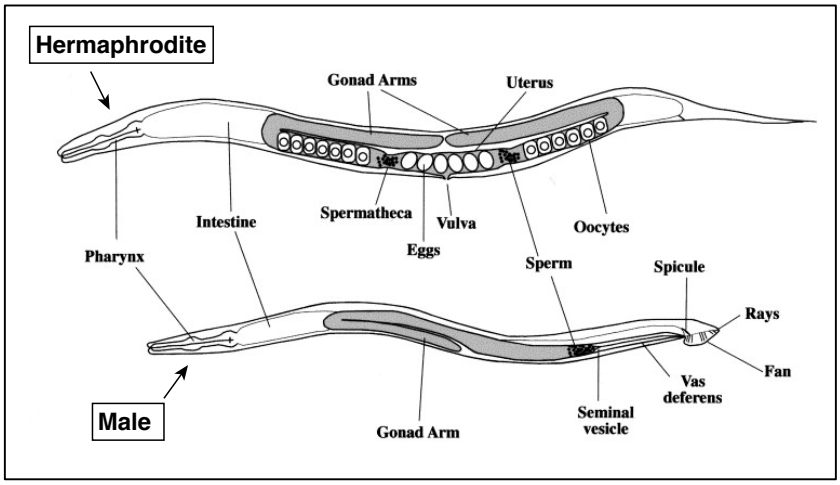


Figure 10 - Morphological differences between hermaphrodites and males (From Hansen and Pilgrim, 1999).

d) Anatomy

As in most nematodes, *C. elegans* has a very simple anatomical plan. It is mainly composed of two concentric tubes separated by a pseudocoelom. The external tube is comprised of the cuticle, the hypodermis, the muscles and some neurons while the internal tube only contains the gut (Sulston and Horvitz 1977). When the worm reaches the adult stage, the pseudocoelom also contains the gonads.

A significant feature of *C. elegans* is its number of cell - 959 in the hermaphrodite and 1031 in males - that is unvarying from one individual to another (Sulston et al. 1983). Each cell has a predefined fate and thanks to the transparency of the animal it has been possible to completely map the lineage of each of its somatic cells from the fertilized egg to the adult, making it an ideal model for the study of developmental biology.

1.3.2 - *Caenorhabditis elegans* as a model organism

C. elegans has several advantages that makes it an easy and affordable model organism. It has a small size (1mm) and can be easily cultivated in Petri dish on a solid growth medium supplemented with *E. coli* bacteria as a source of food. Thanks to their innate resistance, live culture can be kept in dauer stage for several weeks. When needed, liquid cultures can also be carried out. Additionally, stocks can be frozen at -80°C, facilitating the preservation of collections of transgenic and mutant strains.

Finally, with its short generation time of 3-4 days and the size of its progeny, large populations can easily be generated in a short amount of time.

a) Advantages of working with *C. elegans* for the study of genetics

C. elegans exhibits many traits that are significant advantages for the study of genetics, compared to other metazoans:

Thanks to its two modes of reproduction (self-fertilization or mating with a male), it is easy to obtain homozygous animals. Furthermore, the percentage of male worms in a population can be increased by performing a heat-shock treatment, allowing to recover males more easily to perform crosses between different strains.

C. elegans tissues are well differentiated and completely transparent, which makes it possible to express fluorescent markers to study gene expression (level, localization) in living organisms (Chalfie et al. 1994). Additionally, many promoter sequences have already been characterized, including several sequences able to drive tissue-specific expression (Dupuy et al. 2007; Hunt-

Newbury et al. 2007). Hence, allowing for the expression of a given constructs of interest in a specific cellular type.

b) Important contributions made through the study of *C. elegans*

Since the first use of *C. elegans*, lots of important discoveries have been made by studying various aspects of the worms. In recent years, three of those have been recognized with a Nobel Prize:

- In 2002, Sydney Brenner, Robert Horvitz and John Sulston have been awarded with the Nobel Prize in Physiology or Medicine for their discoveries concerning genetic regulation of organ development and programmed cell death (Brenner 1974; Sulston and Horvitz 1977).

- In 2006, Andrew Fire and Craig Mello were also awarded with a Nobel Prize for their discovery of RNA interference - gene silencing by double-stranded RNA (Fire et al. 1998).

- In 2008, Osamu Shimomura, Martin Chalfie and Roger Tsien have been recipient of the Nobel Prize of chemistry for the discovery and development of the green fluorescent protein, GFP, and its use for the study of gene expression *in vivo* (Chalfie et al. 1994) .

Other notable contributions include:

- Discovery of microRNAs: In 1993, the Ambros group discovered two genes (*lin-4* and *let-7*) implicated in the timing of the switching of cell fate during development (Lee et al. 1993). To their surprise, these two genes were revealed to be small non-coding RNAs - termed micro RNAs. This class of non-coding RNAs is able to regulate gene expression by interacting with the RNA sequence of their target. This interaction then leads to RNA repression, affecting downstream processes (Ambros 2004).

- Mapping of the nervous system: By following cell lineage and with the use of electron microscopy, it is now possible to reconstruct a complete map of *C. elegans* nervous system, along with the interactions between the different neurons (Hammarlund et al. 2018). A powerful tool for the study of gene expression in neurons.

- Genome sequencing: It was the first metazoan whose genome was fully sequenced (Consortium* 1998)which allowed to determine of all of the genes found in *C. elegans* and paved the way for the emergence of studies focused on the genomics and transcriptomics of the worm.

The inherent characteristics of the nematode *Caenorhabditis elegans* that have been detailed so far, along with some of the important discoveries it permitted, makes it an excellent model organism for the study of various biological processes, and especially for the study of genetics.

1.3.3 - Studying genetics using *Caenorhabidtis elegans*

Genetic screens can be easily set up in the worm for uncovering genes associated with a function (forward genetics). On the contrary, discovery of RNA interference permitted to inhibit gene expression to study the resulting phenotype (reverse genetics).

With its short generation time, hermaphroditic reproduction and ease of use, the roundworm nematode proved to be a very powerful system for the setup of various genetics screens. As such, *C. elegans* researchers benefit of a large toolbox of various genetic manipulation methods well described for both forward and reverse genetics.

a) Forward genetics methods

Forward genetics screenings are usually carried out in the form of genome-wide mutagenesis, using chemical agents (Ethyl methanesulfonate (EMS) or trimethylpsoralen (TMP) with ultraviolet light)- that can induce random mutations in the germline.
After exposition of the parent (P0), recessive mutations are observed in the second-generation (F2) since, for most mutation, the F1 progeny will be heterozygous. Following self-fertilization, the phenotype of interest will be observable in the F2 animals.

In the context of the study of gene expression, it becomes interesting to combine such screens with fluorescent markers. Gene reporters expressing the GFP are used to provide visual indication of phenotypic alterations for the identification of mutants that would be otherwise unrecognizable. Changes in the activity of a gene, in its localization or its splicing status can therefore be assessed in live worms.

b) Reverse genetics methods: RNA interference

The discovery of the mechanism of RNA interference by Fire and Melo in 1998 opened possibilities for setting up reverse genetics screens in *C. elegans*.

The method relies on the administration of long double-stranded RNA molecules to the worms. Once inside the cell, the dsRNAs are fragmented in small molecules of 21-27nt called small interfering RNAs (siRNA). Those molecules can then bind to messenger RNAs by base-

complementarity, which causes inhibition of gene expression through repression of translation. Using the RNA interference pathway, researchers can cause a drastic reduction in gene expression - referred as "gene knockdown", as opposed to "gene knockout" where the gene is completely inactivated - to study the function of genes in live organisms.

Different methods have been established in *C. elegans* for the delivery of dsRNA. It can be performed mechanically by directly injected dsRNAs into the worm, or passively by soaking the worms into a solution of dsRNAs. It is also possible to deliver the molecule by feeding the worms with bacteria expressing the dsRNA. Finally, it is possible to directly expressed dsRNA in vivo using a pair of construct expressing a sense and antisense transgene.

c) Transgenesis

As quickly mentioned above, the study of alternative splicing in *C. elegans* vastly benefited from the use of gene reporters and particularly bi-chromatic reporters allowing to tag different isoforms with different fluorescent markers (FSS).
In the nematode, the alteration of a specific sequence (deletion, point mutations, addition of fluorescent protein marker, etc.) is possible thanks to two methods that have been developed for performing trans-genesis in the nematode: DNA micro-injection and DNA-coated micro particles by bombardment (Mello et al. 1991; Praitis et al. 2001). These two methods allow the delivery of exogenous DNA into the germline of adult hermaphrodites, generating transgenic progeny.

1.3.4 - Current needs for a better understanding of the transcriptome

Even though *C. elegans* was the first metazoan whose genome was fully sequenced, our comprehension of its transcriptome remains limited. Historically, microscopy-based techniques, with the use of gene reporters and fluorescent markers - or immunochemistry methods - were the go-to approaches for the study of different gene's expression (activity, localization, timing). However, while several technical improvements have allowed the scaling up of this strategy to approach genome wide exploration of gene expression (Bao et al. 2006; Dupuy et al. 2007; Hunt-Newbury et al. 2007; Gerstein et al. 2010). However, those high-throughput techniques usually ignored alternative splicing for which custom reporters need be generated for each case (Kuroyanagi et al. 2006, 2010; Calarco et al. 2007; Watabe et al. 2018).

The construction of splice variant specific gene's reporter is time consuming and is heavily based on current annotations. Yet, some of the available annotations have only been predicted by bioinformatics analysis. In the context of the study of specific mechanisms like alternative splicing, it can be difficult for a researcher to estimate the relevance of a functional isoform, which in turn

can lead to unnecessary efforts being spent for the study of an isoform that is expressed too weakly to be efficiently detected *in vivo*.

With the emergence of cost-effective sequencing methods, we have started to study gene expression in a transcriptome-wide manner and our comprehension of *C. elegans* transcriptome has first been improved with the study of expressed sequence tag (EST) and cDNA-based sequencing libraries. In the last 20 years, a certain number of studies have aimed at refining our knowledge of the worm transcriptome. Some of those methods are detailed below.

a) Investigation of *C. elegans* transcriptome using RNA-seq methods

In 2011, a genome-wide analysis was performed in *C. elegans* to study the alternative splicing (Ramani et al. 2011). The authors combined high-throughput RNA-seq with microarray profiling to study alternative splicing across development. Their results allowed to uncover thousands of new alternative splicing events and hundreds of isoforms that are found expressed differentially during development. Furthermore, this study helped in identifying candidate cis-elements that play a role in the regulation of AS events. At the time of the publication, this work provided the most complete set of splice variants and served as a basis for the study of splice isoforms *in vivo* by the use of fluorescent reporters.

In 2017, a study combined the power of RNA-seq and *in vivo* binding assay for studying how four different splicing factors (SFs) containing well described RNA Recognition Motifs - ASD-1, FOX-1, MEC-8 and EXC-7 - combines to regulates splicing (Tan and Fraser 2017). In this study, the authors were able to report those four SFs regulates many of the same targets and that combinatorial interactions between them affect both individual splicing events and organism-level phenotypes. Taken together, their findings permitted to show that precise splice variant are often the result from the regulation of multiple SFs

The same year, a meta-analysis of alternative exon usage in *C. elegans* was completed by our group (Tourasse et al. 2017). By pooling all of the available RNA-seq dataset, we were able to produce a curation method for discriminating between robust splicing events and biological noise, enabling us to generate measurements of alternative splicing for each of the genes of *C. elegans* (**Figure 11**). Additionally, thanks to the increased detection power of their method, this work also uncovered evidences of trans-splicing for ~3 000 new genes, suggesting the mechanism is more pervasive than previously thought.

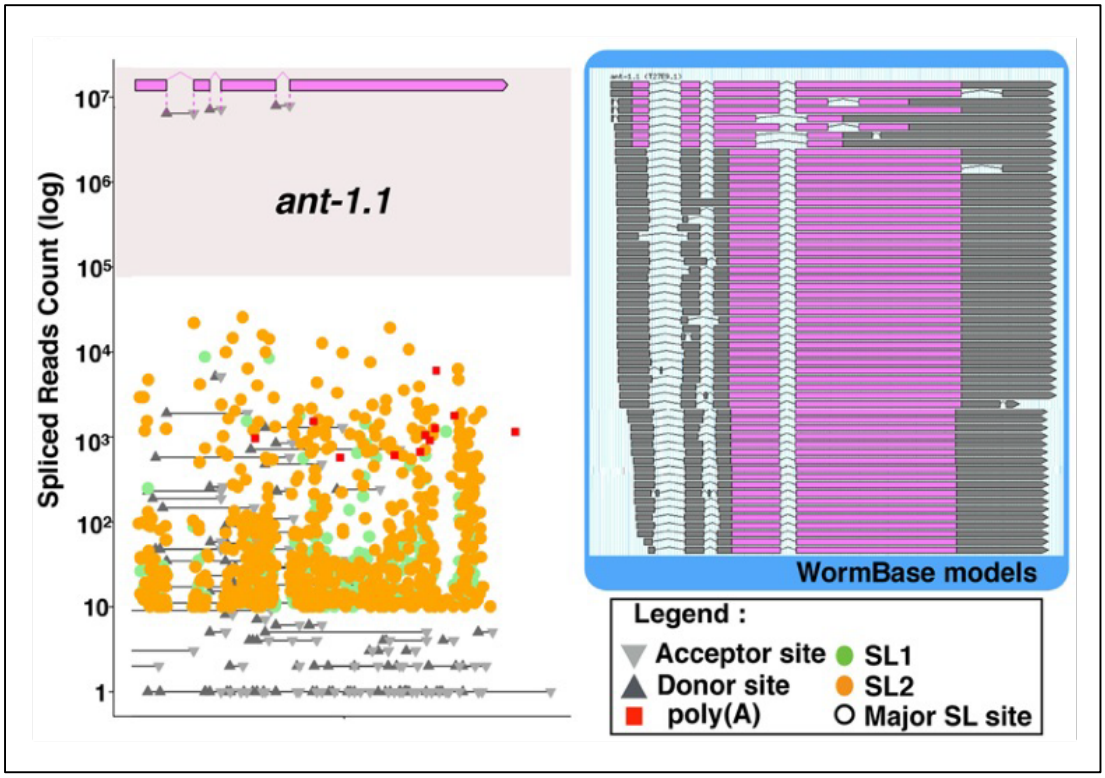


Figure 11 - Quantitative visualisation of relative splice-sites usage. For each gene, exon-junctions are quantitatively measured, allowing to identify robust splicing events from biological noise. This can then be compared to current gene's annotation to determine most prevalent isoforms. Poly(A) and trans-splicing sites are also annotated. **(From Tourasse *et al*, 2017)**

In 2018, the Korswagen group developed a method for performing spatial-transcriptomics in the nematode worm (Ebbing et al. 2018) and generated the first high-resolution, anteroposterior gene expression maps of *C. elegans* males and hermaphrodites (**Figure 12**). To do so, they used RNA tomography, a method combining classical histological sectioning of tissues with the high sensitivity of RNA-seq.

Due to the invariant anatomy of the animal, spatial transcriptomics is an especially powerful tool in *C. elegans* as expression maps between different animals can be precisely compared to determine spatial differences in gene expression. In this study, the method was used for identifying differences between males and hermaphrodites but the generation of maps from wild-type animal will allow to perform comparisons with mutants, offering new possibilities for the study of gene expression.

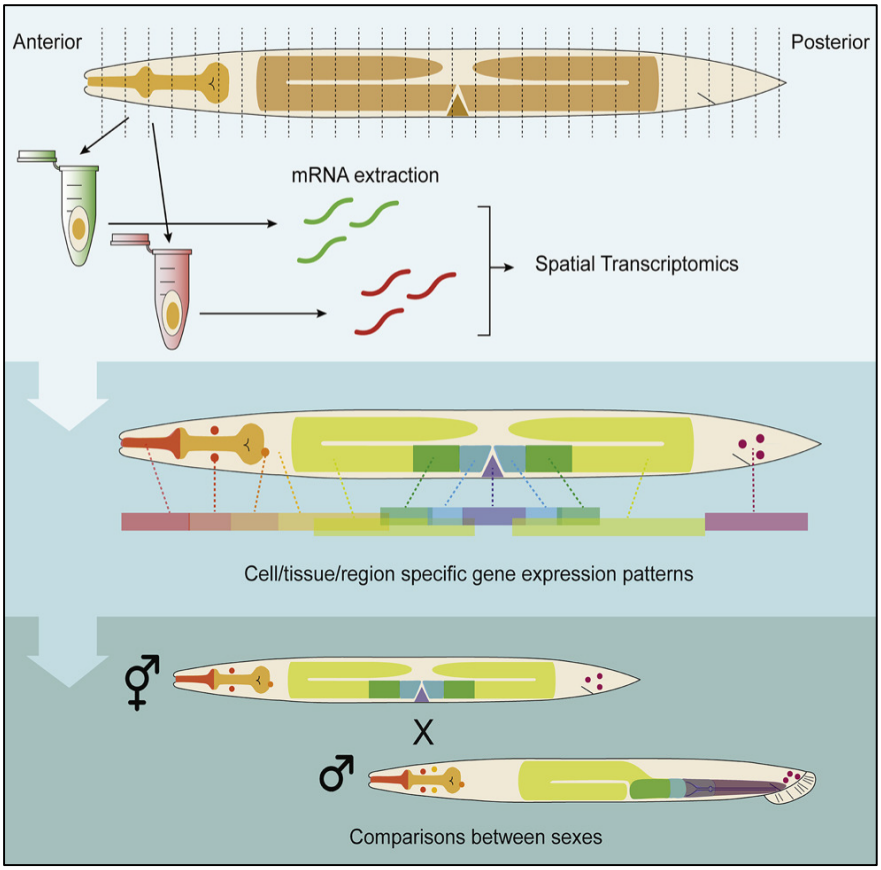


Figure 12 - Generation of an anteroposterior gene expression map of *C. elegans* by RNA tomography and single-cell sequencing. Thin antero-posterior slices of C. elegans are analysis by single-cell sequencing methods to generate spatial transcriptomics maps. **(From Ebbing *et al*, 2018)**

b) Single-cell sequencing: increased resolution for the study of gene expression

Recent years have also seen the development of single-cell sequencing methods (scRNA-seq). The methods rely on the dissociations of the cells of an organism and uses chemical reactions to specifically tag the RNAs coming from the same cell. After sequencing, it is then possible to identify expression patterns that are specific to subpopulations of cells, allowing to observe differences in gene expression at a level we could not reach before. However, this is an expensive method, and it requires the establishment of complex protocols for the isolation of cells.

Nonetheless, *C. elegans* stands as an ideal model for the conduction of scRNA-seq studies due to its completely mapped cell lineage.

In 2016, a first scRNA-seq study was performed on early embryos (up to 16-cell stage) (Tintori et al. 2016). However, the method required manual dissection of each of the embryos, making it hardly scalable for the study of more advanced stages. Combined with cell-lineage map, they were

able to observe transcriptome differences as cells progressively diverged in fate and morphology. They also identified distinct signatures of cell-specificity, along with uncovering the involvement of new genes whose role in developmental program were previously underestimated.

A more recent study was also able to push the method one-step further. The team generated cell suspensions from L2 larvae and used a combinatorial indexing method for uniquely barcoding transcriptomes for a large number of cells (Cao et al. 2017). Based on differences at the level of gene expression and with the identification of cell-specific markers, they were able to cluster resembling transcriptomes together and to assign them to specific cell-types (**Figure 13 - panels B and C**).

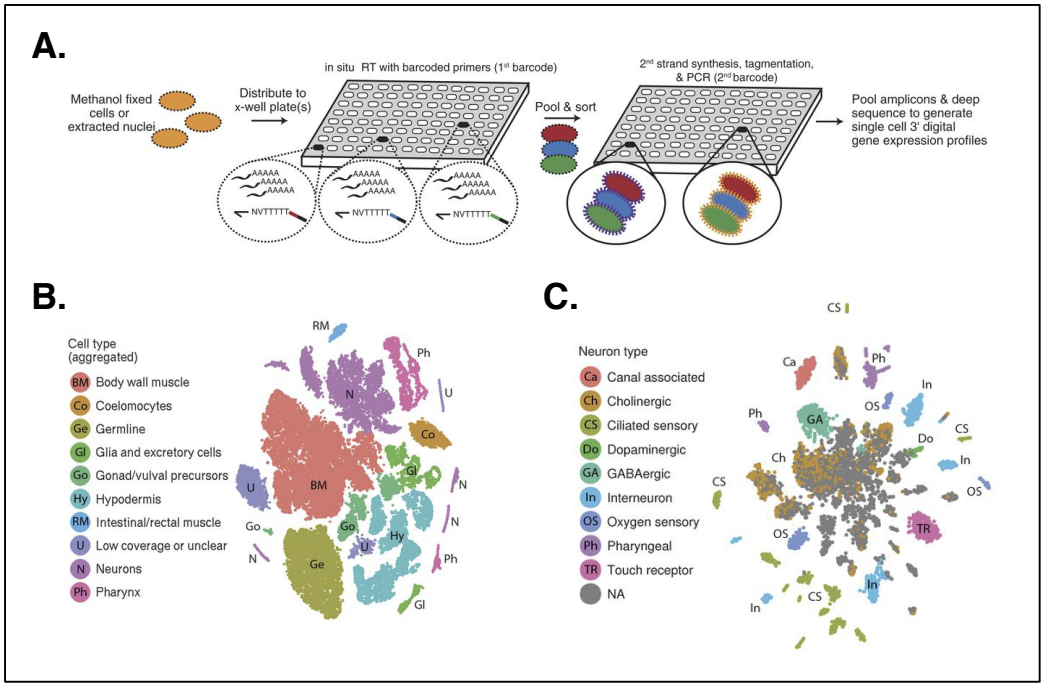


Figure 13 - Single-cell RNA sequencing in nematodes. A) Description of the experimental protocol for uniquely barcoding transcriptomes. **B)** t-SNE visualization of the high-level cell types**. C)** Re-clustering of neurons (from panel B) into sub-neuronal cell types by t-SNE analysis. **(From Cao *et al*, 2017)**

c) Nanopore sequencing permits new approaches for transcriptomics studies

Nowadays, most sequencing studies are being performed with Illumina technology, but this technology is proving to be limited to fully characterize one's transcriptome. PCR-based sequencing methods are known to introduce amplification bias which can negatively affect the overall distribution of mRNAs detected in one's experiment. Furthermore, short-reads are not suited to accurately predict the frequency of isoforms derived from trans-splicing or alternative splicing events, nor to explore poly(A) tail length, 3'-UTR isoforms and RNA base modifications.

In order to address some of its shortcomings, a new generation of technologies has started to emerge. In the next section is given an overview of the main sequencing technologies that have been introduced over the last 40 years. The evolution of the techniques, along with their inherent advantages and disadvantages will be detailed in the next section, as well as why we are now considering the emergence of third-generation technologies, like nanopore sequencing, as an efficient tool for further characterizing *C. elegans* transcriptome.

In the recent years, we have seen the emergence of nanopore sequencing technology for the study of transcriptomics.

In 2020, a transcriptome-wide analysis using nanopore-based direct-RNA sequencing was performed in *C. elegans* (Roach et al. 2020). The authors have demonstrated that full-length reads could be used for the easy detection of novel splice isoforms. However, due to how direct-RNA sequencing libraries are generated, the technique remains unadapted for the characterization of trans-splicing events that takes place at the 5' extremity of the messenger RNAs.

Another study recently submitted on the bioRxiv preprint server took advantage of nanopore technology for carrying complex sequencing protocols without requiring the use of external facility. The authors have established a new method for the study of genome-wide transcription patterns (Gómez-Saldivar et al. 2020). To do so, they used a Dam methyltransferase fused to a RNA polymerase subunit to create transcriptional footprints via methyl marks on the DNA of transcribed genes. By driving Dam fusion expression in specific tissues of interest, they were able to identify genes that are actively transcribed in a pair of XXX neuroendocrine cells (corresponding to 0.2% of the cell content of *C. elegans*), which establish this method as a valid approach for the generation of tissue-specific transcriptional profiles without requiring the use of cell sorting or RNA tagging.

In the next section, the evolution of sequencing technologies over the last 40 years will be addressed. An overview of the main sequencing technologies, along with their inherent advantages and disadvantages will be given.

1.4 - Overview of sequencing technologies

in 1953, with the work of Rosalind Franklin and Maurice Wilkins in crystallography, Francis Crick and James Watson were able to uncover the double-helix structure of the DNA molecule (WATSON and CRICK 1953). This discovery greatly improved our understanding of how genetic information is stored and passed on to next generations. Additionally, the same year, Frederik Sanger was able to sequence the first biological molecule: the bovine insulin (Sanger and Tuppy 1951).

Both discoveries led scientists to question the link between the order of nucleic acids inside the DNA molecule, and the order in which amino acids are being organized within polypeptide chains. Therefore, establishing a need for finding methods which would allow researchers to decipher the content of any genetic material.

1.4.1 - 1st generation sequencing

If early attempts at sequencing nucleic acids (both RNA and DNA) started in the 60s, it only became possible to reliably and efficiently sequence DNA-based genomes with the development of two founding methods in the mid-70s: The Sanger sequencing method and the Maxam-Gilbert sequencing method.

For the establishment of those protocols, both Gilbert and Sanger were awarded with the 1980 Nobel Prize in chemistry for "their contributions concerning the determination of base sequences in nucleic acids", highlighting the ground-breaking nature of their work. This discovery paved the way for the emergence, a few decades later, of a new field in biology that we are now calling "genomics".

a) Maxam-Gilbert sequencing

In 1977, Allan Maxam and Walter Gilbert published a method for sequencing DNA, based on the chemical degradation of a single strand of DNA (Maxam and Gilbert 1977).

Following DNA extraction, the 5' ends of the molecule are labelled with a radioactive isotope of phosphorus (^{32}P). Double-stranded DNA is then denatured in single-stranded DNA by heat and four different chemical reactions are then conducted in parallel, each one specifically designed to degrade one or two nucleic acids:

- Guanine and Adenine (G+A): Formic acid is used to selectively remove purines.
- Guanine (G): Dimethyl sulfate is used to methylate purines. However, methylated adenines being less stables than methylated guanines, this causes guanine to be preferentially cleaved.

- Thymine and Cysteine (T+C): A hydrazine treatment causes both pyrimidines bases to be hydrolyzed.
- Cysteine (C): With the addition of NaCl to the hydrazine reaction, thymines are no longer affected, leaving only cysteine to be hydrolyzed.

In addition to each treatment, a piperidine treatment at 90°C is carried out in order to cleave the modified bases. The concentration of each chemical is calculated so that only one base per molecule is generally affected, producing a library of fragments of every size possible.
The different reactions are then run on an electrophoresis gel (one per lane) for size-separation and fragments location on the gel are revealed using X-rays sensitive films. By comparing the absence or presence of a band in lane G+A and lane A (or lane G+C and lane C), it is possible to determine which of the two base is present at a given position, allowing to decipher the full sequence of the molecule.

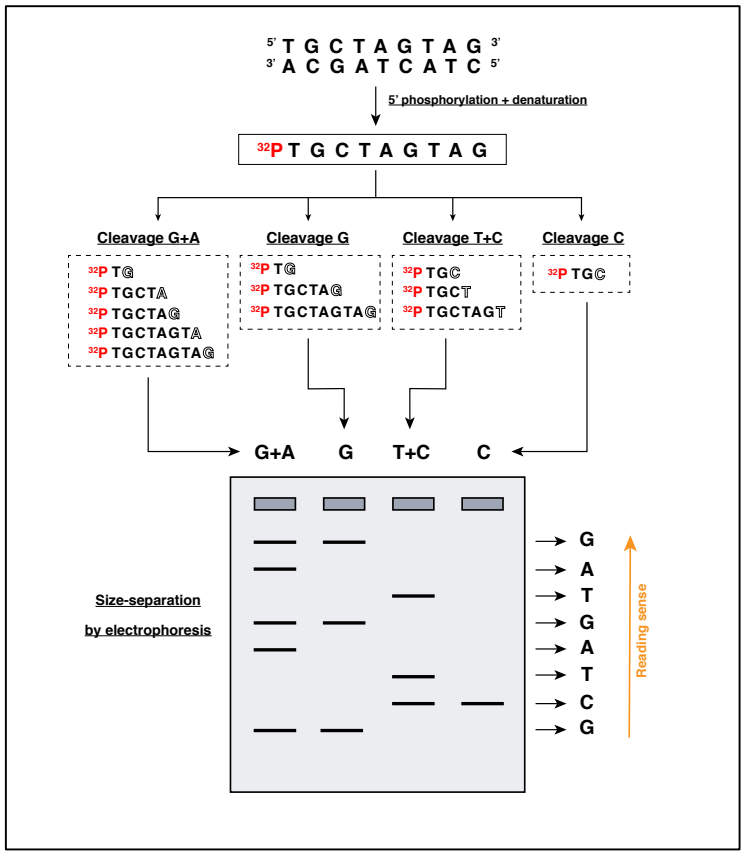


Figure 14 - Maxam-Gilbert sequencing method. A DNA strand is subjected to different treatments leading to cleavage of specific bases. The resulting fragments are size-separated on an agarose gel. G+A and G lanes are compared together to determine the exact base, and the same process is applied to lanes T+C and C. The final sequence is then read from bottom to top.

However, because it was not possible to efficiently automate the procedure, and because of its extensive use of radioactive and hazardous chemicals, Maxam-Gilbert sequencing was rapidly deprecated in favour of another method: The Sanger sequencing.

b) Sanger sequencing

This method was developed by Frederik Sanger and its group in 1977, the same year as Maxam-Gilbert method. The method's core-principle revolves around the idea of coupling the natural process of DNA replication with a chemically-induced chain-terminating reaction (Sanger et al. 1977). This is produced by the use of modified di-deoxynucleotides triphosphates (ddNTPs) lacking a 3'OH group. This causes the elongation process to stop since a new phosphodiester bond cannot be formed anymore by the polymerase, preventing the addition of the next nucleotide in the newly synthesized strand.

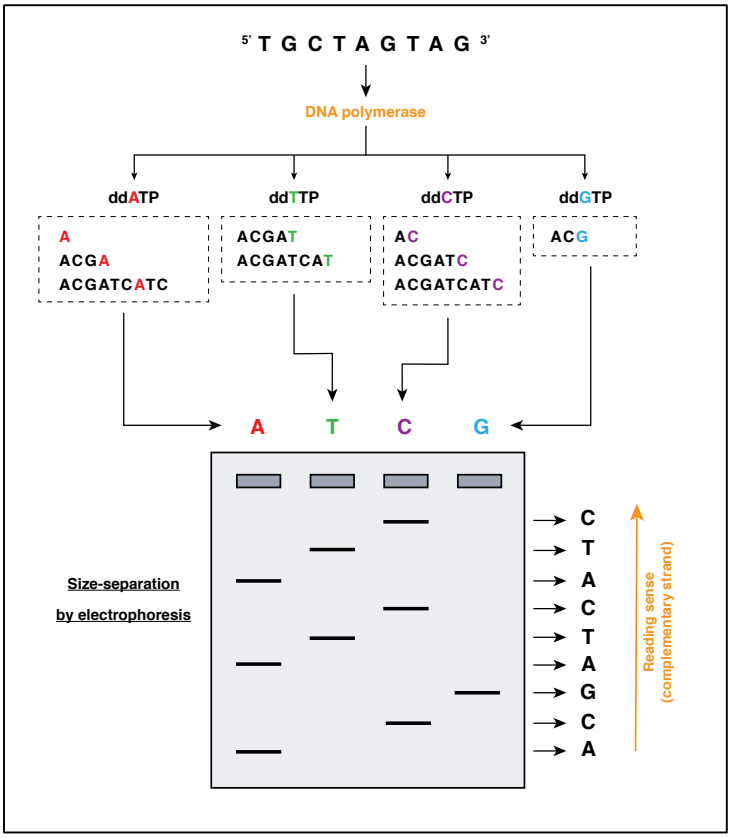


Figure 15 - Sanger sequencing method. A DNA strand is replicated using chain-terminating reactions. The obtained libraries of fragments are then size-separated by gel electrophoresis on a polyacrylamide gel and revealed by X-rays, allowing to read the complementary sequence of the initial molecule.

Four reactions are set up in parallel. In each tube, single-stranded DNA, a primer, a DNA polymerase and a mixture of the four deoxynucleotides triphosphates (dNTPs) is added to allow DNA synthesis. Additionally, each tube gets a different radiolabelled ddNTP - either ddATP, ddTTP, ddGTP or ddCTP. Between 30 and 40 rounds of DNA extension are repeated and each reaction is then deposited onto a polyacrylamide gel for size separation. X-rays are used to reveal the position of fragments that incorporated the modified ddNTP, allowing to extrapolate the sequence of the complementary strand. Finally, this sequence is reverse-complemented.

Thanks to various technological improvement over the years, it has been possible to fully automate the protocol and supplementary advances in the field of microfluidic made it possible to miniaturize the whole system by performing electrophoresis inside small capillaries. In particular, fluorescent dyes have now replaced radiolabelling, allowing to detect the incorporation of a specific ddNTP through the use of lasers, which greatly contributed to reduce both the time and the cost of Sanger sequencing, while also improving its yield and its ability to accurately sequence long sequences.

Today, despite the explosion of Next Generation Sequencing (NGS), Sanger sequencing remains a "gold standard" with 99.99% of accuracy and is still routinely used in laboratories for projects that do not require very high throughput, such as verifying PCR products or plasmids constructs. Moreover, since reads can range up to 900 bases, it is also an excellent method to use when short-reads technologies fail to give an accurate overview of a genomic region.

1.4.2 - 2nd generation sequencing: Next Generation sequencing

While Sanger sequencing allowed for the completion of many different genome projects in the decades that followed its publication, it took 13 years and the combined efforts of many different research teams across the world to fully sequence the ~3,2 billion bases that compose the human genome. And with the growing demand for sequencing bigger genomes, or sequencing different individuals from the same species, technologies with a higher throughput and a lower cost started to emerge.

Those new platforms, termed as Next-Generation Sequencing (NGS) technologies, can be classified in two major categories: platforms that built upon the principles laid out by Sanger sequencing, who perform sequencing by synthesis (SBS), and the ones that took advantage of the double-helix nature of DNA to carry out sequencing by hybridization (SBH).

a) Sequencing by synthesis

SBS technologies refers to a group of methods that performs sequencing through the use of a polymerase or a ligase. Most of these new platforms are inspired by Sanger's method and were developed in order to address some of its shortcomings. The most notable ones are Illumina, Roche 454 and ION Torrent, with Illumina being the most widely used sequencing technology currently.

The Illumina method was developed by Shankar Balasubramanian and David Klenerman, who had the idea to perform sequencing using fluorescently labelled nucleotide (Bentley et al. 2008). The sequencing procedure occurs in the following step:

1) Library preparation: genomic DNA (or cDNA) is fragmented into small sequences (200-500bp) and adapter sequences are added at each end through a ligation step. The adapter sequence is composed of three different regions: one region complementary to the oligonucleotides on the surface of the flow-cell, one region acting as a barcode and one region complementary to the sequencing primer. The fragments with adapters are then amplified by PCR.

2) immobilization: single-stranded fragments are randomly hybridized to the oligonucleotides on the surface of the flow-cell and a polymerase is used for synthesis of the complementary strand. The resulting dsDNA is then denatured and the original strand is washed away, leaving only the newly synthesized strand immobilized onto the surface.

3) Bridge amplification: The ssDNA fold over and bind to the nearest complementary flow-cell adapter. A polymerase binds to the newly hybridized end and synthesize the reverse strand. Both strands are finally separated by denaturation, leaving two complementary ssDNAs attached to the surface.

4) Clonal amplification: The previous step is repeated 30 times, creating a cluster of forward and reverse strands originating from the same library fragment. Finally, reverse strands are cut off and washed away, leaving only forward strands.

5) Sequencing cycles: A sequencing primer is hybridized onto its complementary region on the adapter region of the ssDNA strand, allowing for sequencing by synthesis. During this step, fluorescently-labelled nucleotides are used, which causes the reaction to stop after the addition of each nucleotide because of the fluorophore acting as a blocking group. A laser is used in order to excite the fluorophore and the emitted light is captured, allowing to determine which base was added. Once each cluster has been recorded, the fluorophores are removed by a specific chemical compound, leaving the possibility to add another nucleotide and the next cycle can begin.

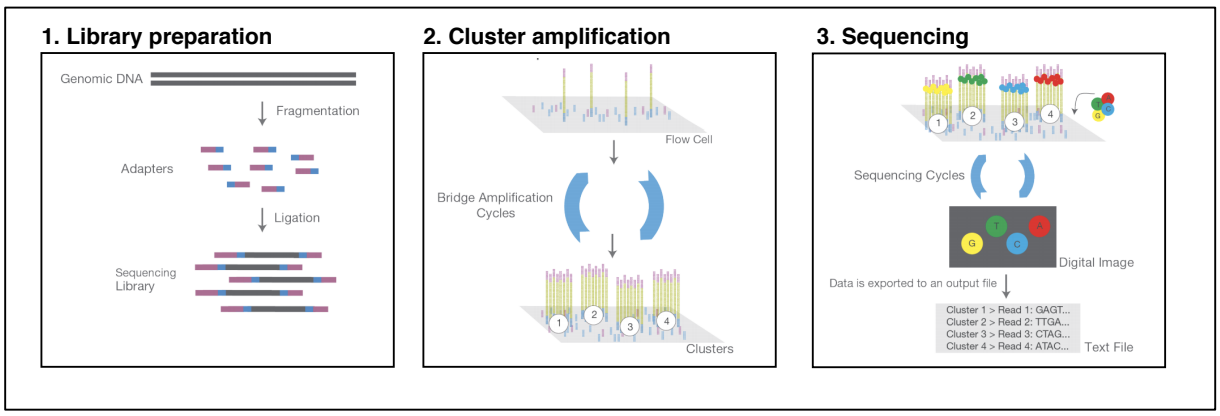


Figure 16 - Illumina sequencing method (Adapted from Illumina.com).

One of Illumina's main advantage over other competing technologies is its relatively low price. While it still cost about $50,000 to fully sequence a human genome in 2007, the Illumina sequencer of 2015 made it possible to reach the record-price of $1,000 per human genome. Despite the considerable cost of the machine, such feat lead to the emergence of more research projects focused on the study of genetics variants within specific populations, such as the "1000 Genomes Project" that had already started in 2008 and used Illumina technology in its last phase.

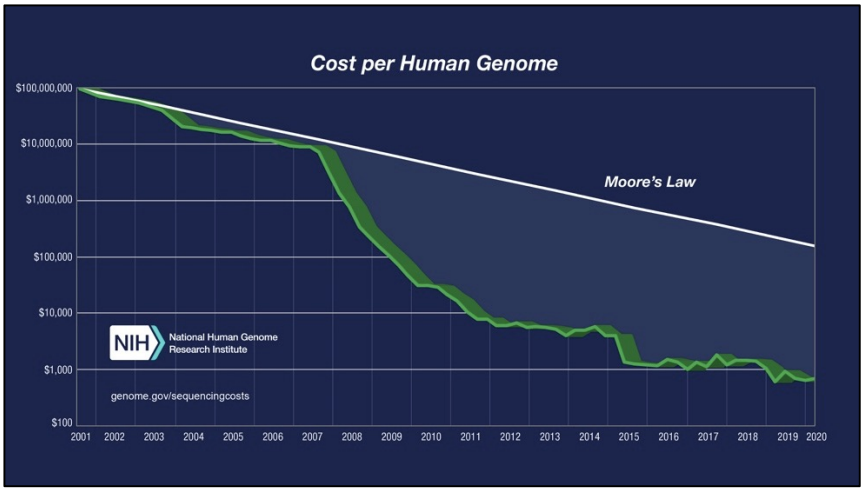


Figure 17 - Sequencing cost per Human genome from 2001 to 2020 (From the National Human Genome Research Institute).

Another key aspect that helped popularize Illumina sequencing is its wide range of application. Over the years, different sequencers have been commercialized, and their differences in term of maximum output, read length and run time makes them more or less adapted to some applications.

On the other hand, Illumina's main disadvantage comes from its inability to produce reads longer than 100bp, making it complicated to accurately sequence - and reconstruct - regions exhibiting highly repeated sequences such as short tandem repeats.
Additionally, Illumina is not always suited for some specific transcriptome analysis. Short-reads are not adapted to determine which of the combination of splicing events observed are coming from the same transcript. Another shortcoming of SBS technologies when it comes to transcriptome analysis is the use of PCR amplification during library preparation which can ultimately affects isoforms ratios observed in the data generated although the use of Unique Molecule Identifiers has recently become more prevalent to circumvent this issue).

b) Sequencing by hybridization

With the increasing number of genome projects completed in the last two decades, reference genomes for most model organisms have become readily available to the scientific community, which laid groundwork for the development of indirect - but cost-efficient - sequencing methods.

The technique's core principles take advantage of the ability of a given DNA strand to hybridize to its complementary strand. A hydrogen bond is formed between complementary nucleotides, which allow to discriminate less-bonding non-specific hybridizations by repeatedly washing them out. In combinations with fluorescently labelled samples, it is possible to determine which DNA sequence (probes) are matching with the sample tested and therefore to know their DNA sequence. This system is traditionally sold as solid-state DNA chips (microarrays) that contains a huge collection of synthetic oligonucleotides bonded to the surface of the chip where each different spot represents a set of specific sequences (feature) to be tested.
Gene expression can also be tested by using a collection of cDNAs sequences retro-transcribed from mRNAs.

Due to the nature of such systems, SBH is now mainly used in the context of diagnostics (Jeffreys et al. 1985). It allows for the identification of known single-nucleotide polymorphism (SNPs) in genes that have been associated with genetic diseases, or for identifying chromosomal abnormalities. Those can range from structural alterations of DNA regions (deletion, duplication or translocation of whole sections of the genome) to copy-number variations (CNVs) as in Down's syndrome.

Nonetheless, because of their very low cost, DNA microarrays are still used in large-scale genetics studies, such as in genome-wide association study (GWAS), where thousands of individuals with a particular phenotype can be genotyped at once which can help to link the impact of a genetics variants with the emergence of a given phenotype (Haines et al. 2005).

1.4.3 - 3rd generation: Full-length sequencing

To overcome the inherent limitations of short-read technologies, it was interesting to develop new solutions that would allow researchers to better understand the structure and the dynamics of the genomes and transcriptomes they worked on. This new generations of sequencing technologies are focused on fully sequencing long strands of nucleic acids.

a) Single-molecule real-time sequencing (SMRT)

In 2010, Pacific Biosciences (PacBio) introduced another method for sequencing nucleic acids whose core principles are also inherited from Sanger's method. Like Illumina, it also uses fluorescently labelled dye to perform sequencing by synthesis, however it benefited from recent

technological improvements to perform single-molecule real-time sequencing based on zero-mode waveguides (ZMW) (Zhu and Craighead 2012)

ZMW are nanophotonic visualization chambers consisting in small metallic cylinders (70nm wide) that are illuminated through a glass support. This structure makes it possible to observe individual molecules within a very small detection volume ($20x10^{-21}$ L) while still maintaining a high signal-to-noise ratio (Tanii et al. 2013)

During library preparation, hairpin adapters are ligated on both ends of the fragment that needs to be sequenced. Fragment size can range up from 250bp to more than 25,000bp. The circularization of the molecule permits to sequence several times the same sequence, in both direction (forward and reverse strands), which greatly increase the depth of sequencing.

Once the library is ready, it is deposited onto the ZMWs on the chip. Each chamber contains a single polymerase that is immobilized at the bottom. The polymerase can only bind a single nucleic acid molecule which result in single-molecule sequencing (Eid et al. 2009; Ibach and Brakmann 2009) During sequencing, nucleotides can diffuse in and out of those chambers and when the polymerase encounters the right nucleotides, it takes a few milliseconds to incorporate it to the newly synthesized strand. This process makes it possible to capture the light emitted by the fluorophore attached to the modified dNTP. Finally, the signal recorded in real-time needs to be converted into readable sequences, this process is called "basecalling".

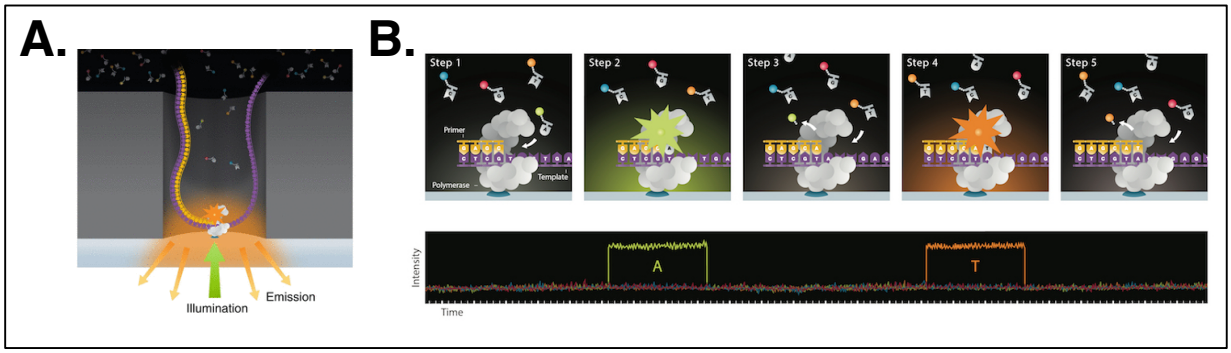


Figure 18 - PacBio sequencing method. **A)** Zero-mode Waveguide containing an immobilized polymerase at the bottom of the well. The emission of light is emitted through the bottom of the well. **B)** Single-molecule real-time sequencing. The incorporation of each modified base generates a specific light that is recorded in real-time by a captor. **(Adapted from Pacbio.com)**

SMRT sequencing has many advantages over previous sequencing technologies.
While a single read can show a very high error-rate (10-15%), most errors are introduced stochastically, making it possible to generate a high-fidelity consensus sequence by increasing the

coverage. Additionally, since it is not necessary to amplify the library by PCR, it reduces the probability to introduce bias in the final data, which is crucial when performing comparative gene expression analysis or for determining isoforms ratios.

Finally, Pacific Bioscience claims it is also possible to detect epigenetics modifications that are present on the nucleic acids, such as methylation marks. This introduce the possibility to study how those changes might affect one's genome or transcriptome without requiring more work or more complex library preparations. Nonetheless, PacBio sequencing is still a very expensive method which yields a relatively low throughput compared to NGS platforms, making it better suited to a set of specific applications that can fully harness the potential of long reads.

b) Nanopore Sequencing

Nanopore-based sequencing was commercialized in 2015 by Oxford Nanopore Technologies and, unlike its direct competitor, its detection principle does not rely on the synthesis of a nucleic acid chain. Instead, it uses the ability of a molecule to affect ionic currents based on the amount of space it takes inside a nanopore. Hence, since each nucleotide has a well-known size, it is possible to infer the presence of a given nucleotide by measuring changes in the current. Therefore, it is possible to reconstruct the sequence of nucleic acid being translocated through a nanopore by measuring changes affecting the current.

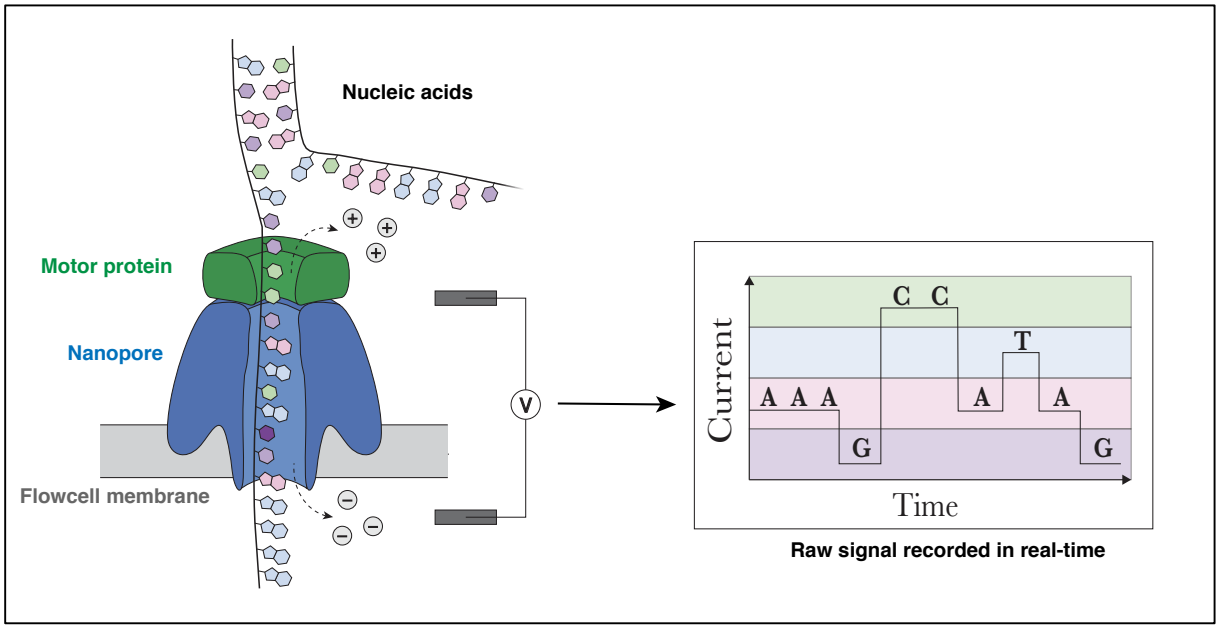


Figure 19 - Schematic of Nanopore sequencing. The nanopore is fixated onto the flowcell membrane. A motor protein (with helicase activity) guides a single strand of nucleic acid inside the pore and, as nucleic acids enter and leave, the occupied space inside the pore is modified inside and affects ionic exchanges. By recording voltage in real time, it is then possible to determine the sequence of the nucleic acid that went through the pore.

Incidentally, this approach permits the sequencing of both DNA and RNA but can also be applied to other applications, such a protein sequencing since it mostly relies on the size of the molecule inside the nanopore.
Additionally, the molecule does not require to be amplified prior to sequencing, allowing to study the molecule in its native state – opening the way for accurately measuring epigenetics modifications present on DNA and RNA molecules – but also preventing the introduction of amplification bias during PCR amplification.

Oxford Nanopore technology has released a certain number of kits suited to different applications. In the context of the study of gene expression, three kits are particularly advised:

- 1D ligation kit for the sequencing of cDNAs amplified by RT-PCR.
- Direct-cDNA kit for the sequencing of cDNAs without PCR amplification
- Direct-RNA kit for the sequencing of RNAs without RT and PCR steps.

Each kit involves a different protocol for the generation of the library, however the last steps are common to most kits and require the addition of specific sequencing adapters at the extremities of the nucleic acids. Those adapters hold a motor protein, that is able to unwind double stranded molecules for translocation of a single strand inside the pore, as well as controlling the speed at which it occurs. In most kits, sequencing is performed from 5' to 3' but both the sense and antisense strand of the molecule can be sequenced (**Figure 20**).

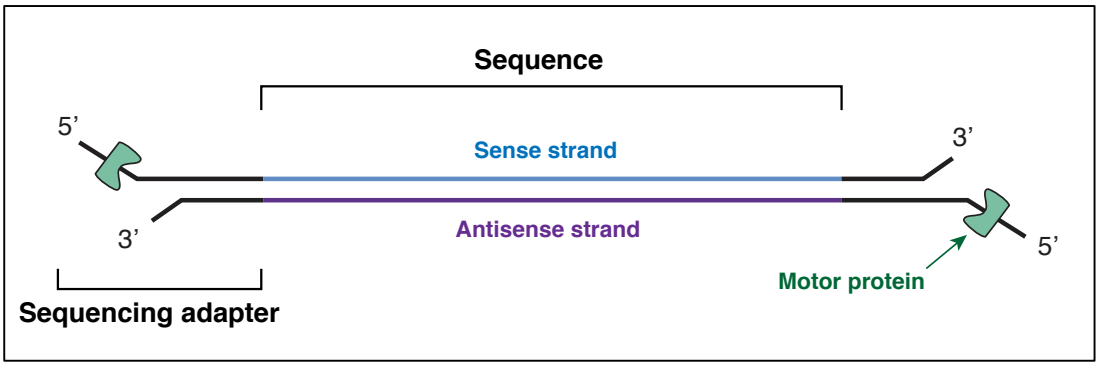


Figure 20 - Schematic of a double-stranded cDNA molecule after library preparation. Sequencing adapters containing a motor protein are added at each extremities of the molecule, allowing to sequence both strand from 5' to 3'.

Thanks to the small size of the MinION sequencer, it is now possible to perform sequencing experiments outside of well-equipped research labs. The sequencer can be brought out on the field to perform experiments as soon as samples are retrieved, making it a powerful tool for the rapid identification of viral pathogens or in environmental studies (Walter et al. 2017).
In 2016, a MinION sequencer was brought on the International Space Station (ISS) and used by astronaut Kate Rubins to perform the first DNA sequencing experiment in space (Wong 2019).

Two years later, astronaut Ricky Arnold directly sequenced RNA molecules in the same conditions, proving again the effectiveness and the versatility of the technology.

Another advantage of ONT's sequencers is the possibility to "sequence as much as we need". Statistics regarding the experiment are reported in real-time (number of sequences, fragment size, etc.) and the user can decide to stop the run when enough data have been acquired. On the other hand, if more data are needed, the acquisition can continue up to 48h per flow-cell. This allow to directly survey the quality of the run, and rapidly stop a bad experiment or extend a successful one. Furthermore, on powerful computers, basecalling of the data can be performed in real-time as they are being acquired, which makes it even easier to control the quality of the run or the presence of a given sequence.

However, it is important to note the high error-rate of nanopore-based technologies which can go up to 10-15%. However, just like SMRT sequencing, it can be greatly reduced by generating a consensus sequence from multiples reads. Furthermore, while the improvement in accuracy is constantly achieved through better kits (better enzymes, newer pores, etc.), it also comes from software updates and the release of better algorithms and better models for the basecalling of the data. Hence, making it possible to re-basecall previous data files in order to improve the quality of the data.

1.4.4 - Summary

Over the last 40 years, sequencing technologies had to evolve in order to adapt to newer needs. First generations technologies like Sanger sequencing were not entirely suited for the sequencing of large genomes, even if it has been possible to greatly improve its efficiency by automatization of the different steps. Second generations technologies, like Illumina, permitted to have a much higher throughput and made sequencing more affordable. Yet, due to the small size of the reads and to the extensive use of PCR amplification, the technology is not convenient for exploring the particularities of a transcriptome. Observed isoforms ratios can be altered during PCR amplification and isoform identification relies on reconstruction algorithms based on fragments spanning exon-junctions. In the context of the study of messenger RNAs, the development of full-length technologies now provides a new way for characterizing the transcriptome of any given species. Despite being less accurate, bioinformatics methods have been developed to improve the quality of the data and the methods do not suffer from the drawbacks of previous generations. Moreover, in the case of transcriptomics studies for which a reference genome is already known the accuracy is sufficient to identify unambiguously the vast majority of long reads. Today, nanopore sequencing technology, due to its affordable price and ease of use, represents a particularly interesting tool for the study of *C. elegans* transcriptome.

Chapter 2

Results

2.1 - RNA sequencing in *C. elegans* with nanopore technology

Following the quantitative RNA-seq meta-analysis of alternative exon usage in *C. elegans* performed by the team of Dr. Dupuy (Tourasse et al. 2017), it was decided to further investigate two key observations that were already discussed in their analysis.

The first observation concerned the relationship between different exons. Some genes present exon-junctions that are expressed at the same level, but it is impossible to determine if those junctions are all coming from the same isoform or, instead, if they come from different isoforms but expressed at the same level. Moreover, junctions showing very different level of expression also raises other questions: are they coming from isoforms differentially expressed, or are they coming from isoforms specifically expressed in certain tissues?

Another question raised by the results of this meta-analysis concerns the mechanism of trans-splicing. Previous studies estimated that 70% of all protein-coding genes of *C. elegans* were subjected to trans-splicing while, in this study, increased detection power allowed to detect that about 86% of all protein-coding genes are being trans-spliced. On top of this, genes for which no trans-splicing was detected are showing very low level of expression, suggesting that the RNA sequencing technologies used for the generation of those datasets were not entirely suited to capture such events. In light of those observations, one can wonder if not all of *C. elegans* genes are subjected to trans-splicing.

To address those different questions at the same time, we decided to perform a transcriptomic analysis with a new sequencing technologies. We opted for nanopore-based technology as the advantage of full-length reads allow us to improve isoform identification and to better quantify the relationship between exon-junctions as well as trans-splicing events. Furthermore, if needed, the possibility to directly sequence mRNAs without needing to amplify the library can allow us to eliminate potential PCR-bias for a more accurate quantification.

2.1.1 - Comparisons between RNA sequencing kits

The first step in setting up RNA sequencing with nanopore technology was to determine which kit is more suited to our needs. Nanopore sequencing being a recent technology on the market, and in constant development, I chose to test three different kits marketed for RNA sequencing:

- 1D ligation: a kit for sequencing cDNAs amplified by RT-PCR
- Direct-RNA: a kit for sequencing poly(A) RNAs
- Direct-cDNA: a kit for sequencing cDNAs without requiring PCR amplification

For each, I generated duplicate experiments and then performed different comparisons between each duplicate and between each kit to determine which kit would be used in the subsequent

experiments. Among the criteria I used, I considered both the quantity and the quality of the data generated, but also other metrics like the amount of input material and the reproducibility of the experiments.

a) Experimental design

Since the different kits have been designed for the sequencing of specific products (RT-PCR products, cDNAs or mRNAs), we had to prepare and process RNAs differently for each application (**Figure 21**).

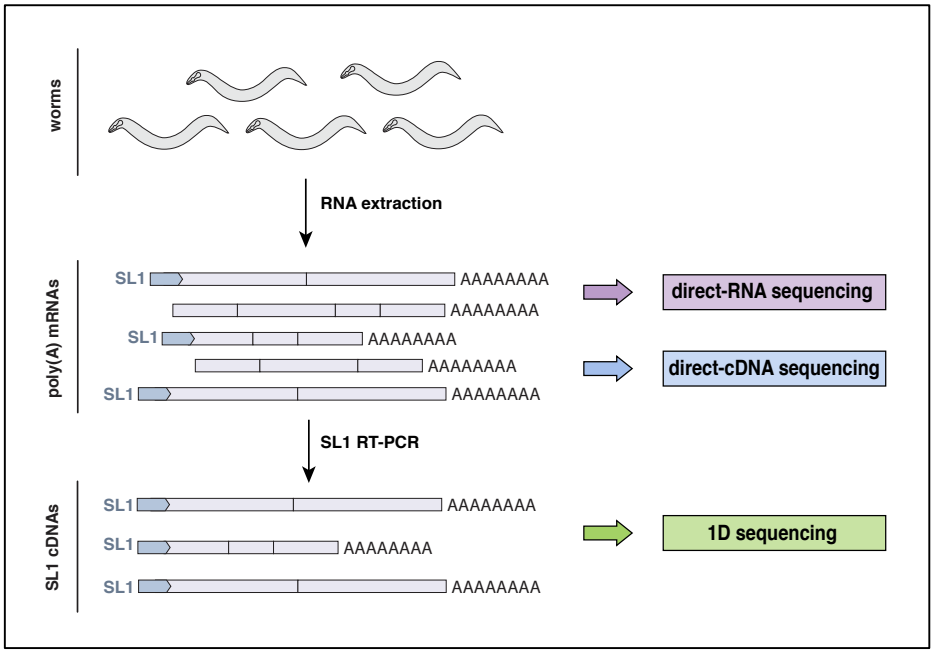


Figure 21 - Experimental approach for the generation of different libraries starting from total RNAs extract. For direct-sequencing, mRNAs are pulled down from total RNAs extractions and then used as is (direct-RNA) or retro-transcribed (direct-cDNA). For 1D sequencing, specific SL1 RNAs are amplified by RT-PCR from the total RNAs extracts.

As a starting material, we used total RNAs extracts from wild-type (WT) worms. Then, we performed RNAs isolations in order to obtain purified poly(A) RNAs that could be used directly for sequencing (direct-RNA kit), or retro-transcribed prior to sequencing (direct-cDNA kit).

For the 1D ligation kit, we opted for an approach that would let us a wide range of cDNAs amplify by RT-PCR (from total RNAs extracts) by using just two primers. This particular amplification was possible thanks to the presence of a splice leader sequence on most of the mRNAs of the nematode. As seen in the introduction, two main splice leader sequences are presents on *C. elegans* mRNAs: SL1, with a unique variant, and SL2 with 11 variants.
Since SL1 is the most commonly found sequence, we decided to use this feature, along with the poly(A) tail, to perform a SL1::poly(A) RT-PCR in order to amplify all of the SL1-mRNAs.

The benefits of this approach allow us to easily amplify a library of cDNAs without needing to modify their 5' extremity first. On top of this, sequencing a SL1 library could allow us to identify genes that had not been detected before due to a low level of expression or confirm the lack of association between the SL1 sequence and specific genes. However, we expect such library would fail to amplify SL2 genes and, therefore, does not provide an accurate picture of the whole set of trans-spliced mRNAs found in the nematodes. In order to fix this problem, complementary RT-PCR using a SL2 primer that is able to amplify mRNAs associated with any SL2 variants could performed or we could attempt multiplexed RT-PCR experiments.

b) Library preparation

For each kit, we prepared libraries according to the recommendations of the manufacturer. An overview of the protocol of each kit is given in the following table:

Sequencing kit	ONT protocol	Recommended input	poly(A) pull-down	retro-transcription	PCR amplification
1D ligation	SQK-LSK108	1ng poly(A) RNAs	NO	YES	SL1::poly(A)
direct-RNA	SQK-RNA001	500ng poly(A) RNAs	YES	NO	NO
direct-cDNA	SQK-DCS108	100ng poly(A) RNAs	YES	YES	NO

Table 1 - ONT recommendations and library preparation. For each kit, the protocol recommended by ONT is noted along with the recommended amount of input RNAs. An overview of the main steps (pull-down, retro-transcription and PCR amplification) is provided.

As the ligation of the sequencing adapters onto the different products is similar between each kit, this will not be presented in this section. I will only provide details related to the amplification of a library of SL1 RNAs (1D sequencing) or related to the isolation of poly(A) RNAs (direct-RNA/direct-cDNA).

Amplification of SL1 RNAs by RT-PCR:

Totals RNAs were isolated from a large population of wild-type adult worms using a phenol-chloroform extraction method followed by ethanol precipitation.
1µg of total RNAs were used for retro-transcription, in a total volume of 20µL. Following this step, 2µL of the reaction was used for PCR amplification of a SL1::poly(A) library. After PCR amplification, the product was purified using a QIAGEN PCR-purification kit and deposited onto a 1% agarose gel for size-separation by electrophoresis. The presence of nucleic acids was revealed by treatment with ethidium bromide followed by exposition to UV lights (**Figure 22.A**).

As expected for the amplification of a diverse population of mRNAs, we obtained a smear, with most products ranging between 500bp and 1200bp. In order to validate this amplification, we compared this result with known sizes of CDS sequences for *C. elegans* mRNAs and due to their similarity, it was decided to sequence this RT-PCR product (**Figure 22.B**).

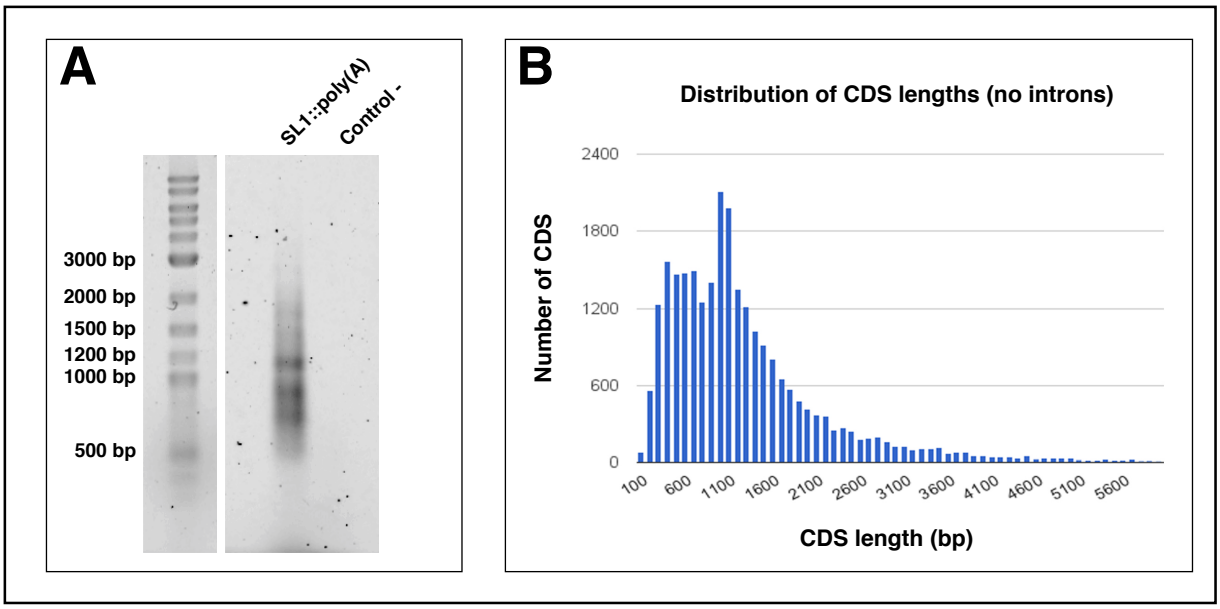


Figure 22 - Generation of a SL1 library. A) RT-PCR amplification **B)** Distribution of CDS lengths.

As recommended by the SQK-LSK108 protocol, 1µg of the purified RT-PCR product was then carried out in the following steps for ligation of the sequencing adapters.

Isolation of poly(A) RNAs:

In order to isolate poly(A) RNAs, we performed a poly(A) pull-down using magnetic beads coated with oligo-d(T) (Dynabeads mRNA purification kit). We incubated 50µg of total RNAs with 10µL of magnetic beads for 8min at room temperature (RT) and the captured RNAs were eluted in 15µL of 10mM Tris-HCL (pH 7.5).

We controlled the isolation by depositing the elution phase onto a 1% agarose gel. 10µL of the supernatant of the isolation (containing non-captured RNAs) was also deposited as a control. Ethidium bromide was added to the agarose to reveal nucleic acids following exposition to UV. As expected from the isolation of RNAs molecules of varying sizes, a smear was observed on the gel (**Figure 23**). However, we can also observe important bands indicating the presence of ribosomal RNAs in our isolation. Nonetheless, we decided to continue forward with this fraction without purifying it more because direct-RNA and direct-cDNA library preparation uses oligo-d(T) primers that are specifically targeting poly(A) RNAs, thus the presence of rRNAs contaminations should not affect the experiments.

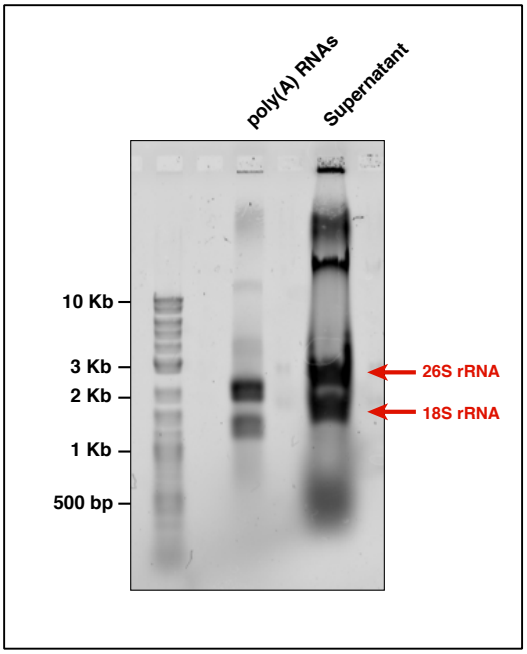


Figure 23 - Isolation of poly(A) RNAs using magnetic beads. 26S and 18S rRNAs contaminations are indicated by red arrows.

c) Quantification of sequencing output:

In order to evaluate the output of each kit, two different measures were performed: the number of reads generated per minute, and the total number of reads generated over time.

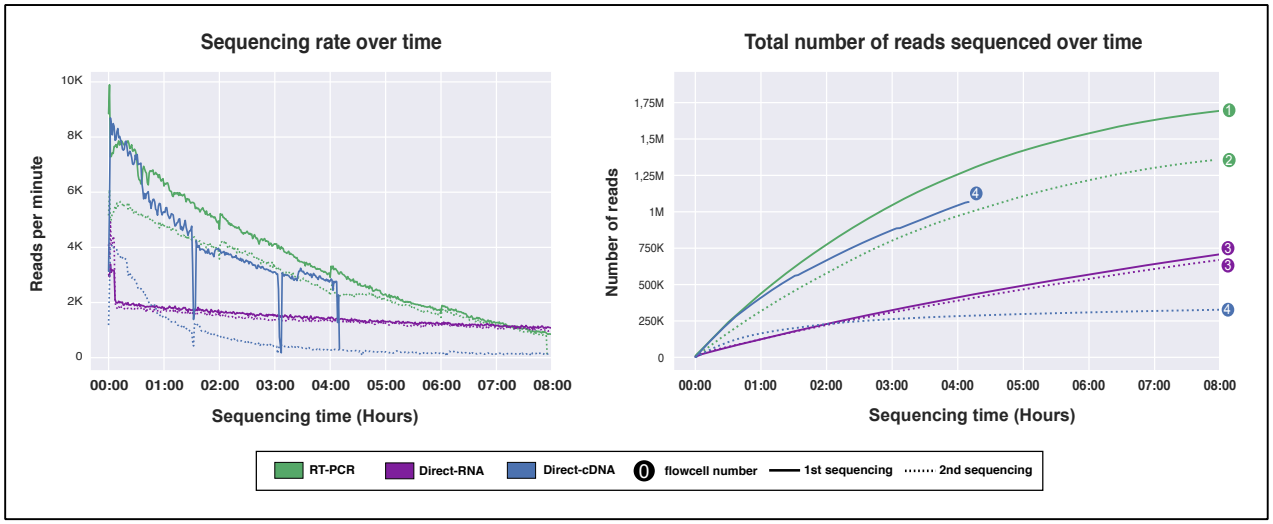


Figure 24 - Sequencing rate over time. The left panel shows the number of reads sequenced per minute and the right panel show the total number of reads sequenced since the beginning of the experiment.

Both DNA-based kits have a much higher throughput than direct-RNA sequencing. This difference is expected since, according to the technical information available on Oxford Nanopore

Technology website, DNA and RNA kits uses different motor protein that have different processivity. For DNA sequencing, the translocation of the strand through the nanopore is performed at 450bp/s whereas for RNA sequencing the translocation is performed at a much lower speed (70bp/s). This difference in processivity is responsible for the lower output produced by direct-RNA kit when compared to 1D ligation or direct-cDNA kits over the same amount of time.

However, we can observe the 2nd direct-cDNA experiment performed poorly compared to the first experiment. It started with a good sequencing rate in the first minutes of the experiment (4000 reads/min) but quickly dropped below the rate of direct-RNA sequencing after just an hour and reached a sequencing rate of less than 100 reads/min after 3h, producing 50% less reads than direct-RNA experiments (**Figure 24** - dotted curve in blue versus purple curves).

Since both direct-cDNA experiments were ran sequentially on the same flowcell (number 4), we decided to look at the activity of the flowcell for each sequencing experiments. Flowcells are composed of 512 channels, arranged in a 16 x 32 grid, and each channel represents a nanopore. By measuring the number of reads sequenced by each channel, we can measure the activity of the flowcell over a given period.

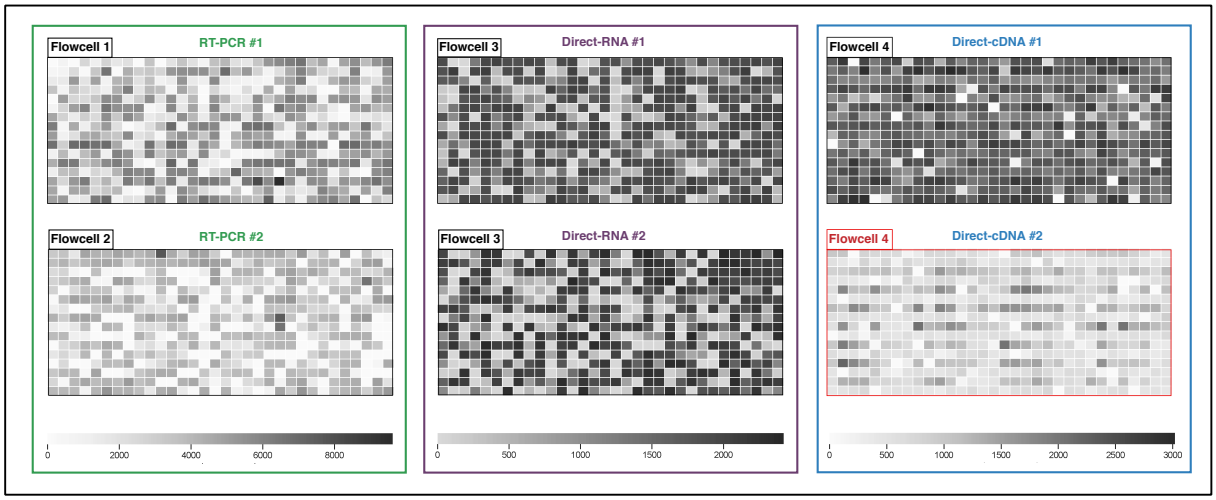


Figure 25 - Flowcell activity during each sequencing run. The grid depicts the different channels of the flowcell, each channel representing one nanopore. The activity of each channel (total number of read sequenced) is depicted as a shade of black and white - white meaning low activity and black meaning high activity. RT-PCR experiments were run on two different flowcells (n°1 and 2). The two direct-RNA experiments were run on the same flowcell (n°3). The two direct-cDNA experiments also ran on the same flowcell (n°4).

The 1st and 2nd RT-PCR experiments were performed on two different flowcells (flowcell n°1 and n°2), but we can observe the general activity of each flowcell was about the same in both experiments (**Figure 25 - left panel**). Both Direct-RNA experiments were run on the same flowcell (n°3), one after the other. We can observe the general activity on the 2nd run was not diminished following sequencing of a first experiment (**Figure 25 - middle panel**).
Similarly, both direct-cDNA experiments were run on the same flowcell (n°4), however we can observe a significant difference between the activity the flowcell in the first experiment and the second (**Figure 25 - right panel**). This difference of activity explains the low output of this experiment. Yet, the flowcell was equally affected for all channels, suggesting that either we had a poor sequencing library (low concentration of cDNAs, poor efficiency during the ligation of sequencing adapters, etc.) or that the nanopore of the flowcells were damaged between the two experiments.

d) Measurement of read quality:

Another key aspect during this analysis was the quality of the reads obtained with each kit. Nanopore technology is known for having high error rates, but since DNA and RNAs kits use different chemistry as well as different algorithms for basecalling, I wanted to see if this would affect the average quality of reads.

During basecalling of a read, each nucleotide is attributed a quality score (Q) known as PHRED score. It estimates the probability of correct identification of each nucleotide by the algorithm. By taking the average Qscore of each read (returned by the basecaller algorithm) of a given dataset, it is possible to determine the mean error-rate using the following formula:
Probability of incorrect basecall: $P = 10^{\left(-\frac{Q}{10}\right)}$

Following this formula, a Qscore of 10 represent an error rate of 10%. A Qscore of 20 represents an error rate of 1% and a Qscore of 30 represents an error rate of 0.1% (see **Figure 26.B**).

For each read, we extracted the mean Qscore and generated a violin plot to show the distribution of this value for each sequencing experiments (**Figure 26.A**). The mean error-rate of each experiment is in the expected range of known error-rate for nanopore technology, with a median value for each dataset situated between 8 and 10 (PHRED score), which represent an error-rate of 10-15% (**Figure 26.B**).

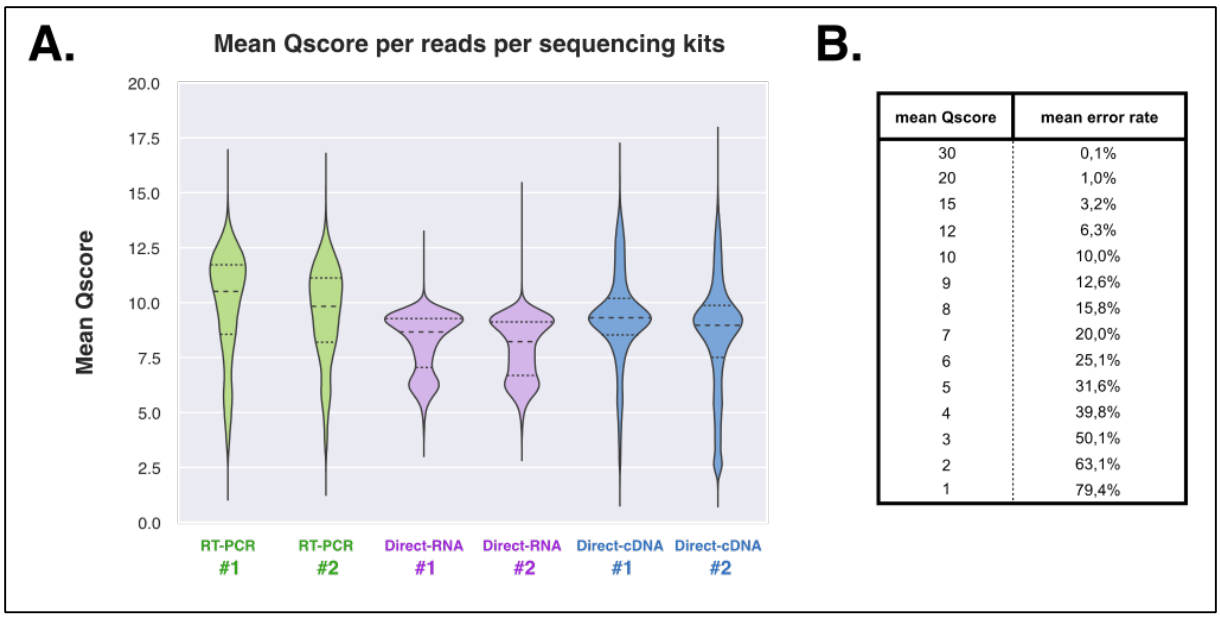


Figure 26 - Analysis of read's quality. A) Distribution of mean Qscore per sequencing experiments. **B)** Correlation between Qscore (PHRED) and error rate.

If the distribution is very similar between each duplicate experiment, we can observe differences between each kit. The sequencing of RT-PCR products with the 1D ligation kit resulted in the best overall quality, with 50% of the reads exhibiting less than 10% error-rate.
Direct-cDNA sequencing performed slightly less well in term of error-rate but shows a more homogeneous distribution as demonstrated by the violin plot. Finally, direct-RNA sequencing produced reads with the lowest overall quality and, unlike in the other experiments, it exhibits a large proportion of reads of reads with a poor Qscore as demonstrated by the lower quartile (lower quartile below 7 in direct-RNA and above 8 in RT-PCR and direct-cDNA).

The quality of the reads was further evaluated by measuring the number of reads that we could map onto *C. elegans* genome and transcriptome.
On average, 89% of reads from the RT-PCR experiments are mapping onto the genome and 87% onto the transcriptome, showing that most of the reads we obtained are from protein coding genes. Concerning direct-cDNA sequencing, we also observe 85% of the reads mapping onto the genome, however only 76% of those were mapped onto protein coding genes. The difference of mapped reads between the genome and transcriptome suggests we have also captured RNAs corresponding to non-coding genes (rRNAs, or other types of non-coding RNAs) and could be confirmed by looking at the region of the genome where those reads aligned.
Finally, only 50 to 60% of direct-RNA reads were mapped onto the genome and transcriptome of *C. elegans*, highlighting the fact that the quality of direct-RNA reads is not as good as the quality of the reads obtained with DNA-based sequencing kits.

Experiment	Total reads	Genome mapped		Transcriptome mapped	
		Reads	Percentage	Reads	Percentage
RT-PCR #1	1 692 943	1 501 474	88,69%	1 479 368	87,38%
RT-PCR #2	1 355 384	1 214 408	89,60%	1 187 766	87,63%
Direct-RNA #1	707 216	462 108	65,34%	433 177	61,25%
Direct-RNA #2	668 887	347 529	51,96%	330 740	49,45%
Direct-cDNA #1	1 067 069	926 986	86,87%	856 411	80,26%
Direct-cDNA #2	327 079	271 210	82,92%	239 075	73,09%

Table 2 - Number of reads mapping onto *C. elegans* genome or transcriptome. "Total reads" corresponds to the number of reads obtained after basecalling. The reads were either mapped onto the genome ("Genome mapped") or the transcriptome ("Transcriptome mapped"). For each experiment, the number of reads mapped and the percentage of total reads it represents is given.

e) Gene expression:

To measure gene expression in each dataset, we looked at read count per gene for each experiment. For each set of duplicate experiment, I plotted the number of reads/gene obtained in one experiment versus the other. The reproducibility between duplicates experiences was evaluated by performing a linear regression analysis and measuring the R-square value (R^2) of the fitting line.

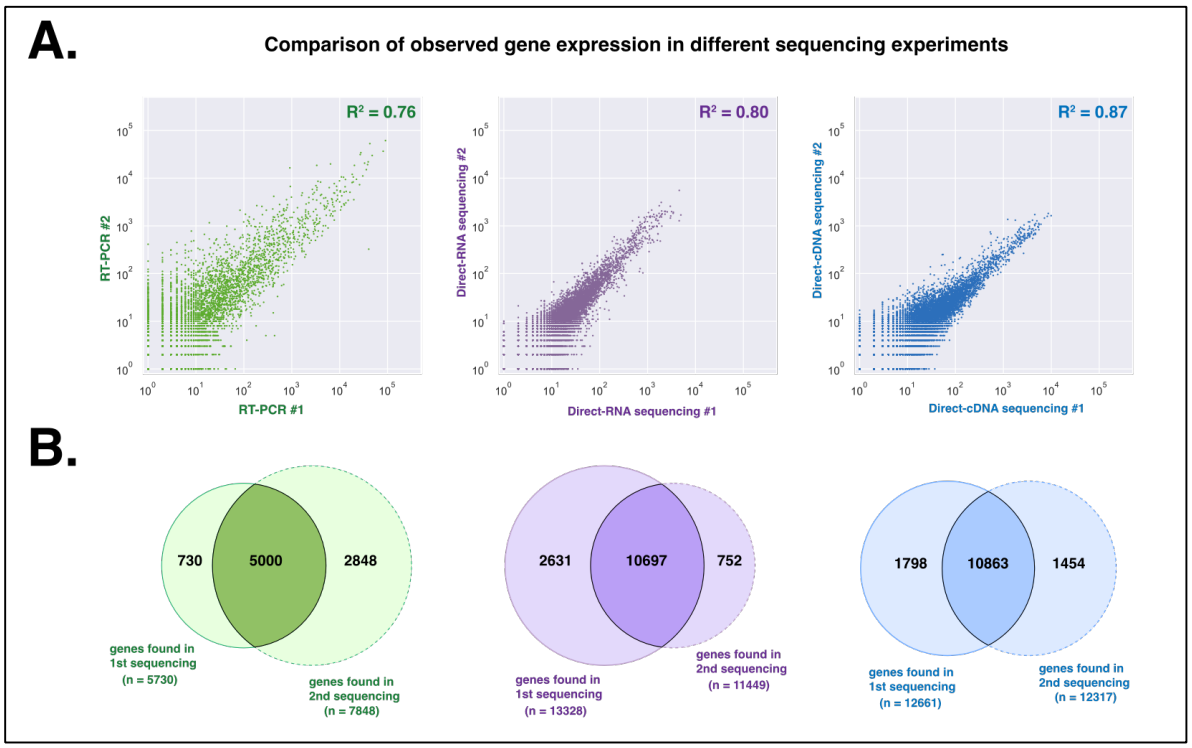


Figure 27 - Reproducibility of the datasets generated with each kit. A) Scatter plots of gene expression (number of reads per gene) between duplicate experiments. **B)** Venn diagrams depicting the number of genes uniquely found in one experiment or found in both.

The result indicates that RT-PCR experiments are less reproducible in term of number of reads/gene (R^2 value of 0.76). From the plot, we can observe genes being detected 10 times more in one experiment versus the other, showing the heterogeneity of the libraries. Direct-RNA and direct-cDNA experiments shows a better reproducibility (R^2 values respectively 0.80 and 0.87).

As an additional measure of the homogeneity of each dataset, I also counted the number of genes found in each one, as well as the number of genes uniquely found in one of the two duplicates (**Figure 27.B**).

The first RT-PCR experiment has about 300K more reads mapping to the transcriptome than the second experiments, yet 730 genes are uniquely found in the first experiment while 2848 genes are uniquely found in the second, suggesting that a greater number of reads does not equal to a better detection power when using this kit. Additionally, when pooled together, those two datasets only manage to capture 7 992 genes despite ~2.5M reads mapping to *C. elegans* transcriptome.

In direct-RNA experiments, we detected a total of 10 697 genes that are found in both experiments, with an addition of 2631 unique genes only detected in the first experiment and 731 in the second, bringing the total to 14 080 genes detected with only 760K reads mapping to the transcriptome.

Direct-cDNA experiments performed similarly as direct-RNA experiments, with 10 863 genes being detected in either experiments and respectively 1 798 and 1 454 genes uniquely found in the first and second experiment. By combining the two datasets, we detected 14 115 genes for a total of ~1,1M reads mapping to the transcriptome.

We then pooled together the data obtained from the replicate experiments to compare gene sets detected with each kit.

From the upset plot in **Figure 28**, we can see that only 200 genes are uniquely found in RT-PCR experiments while the rest is found in either direct-RNA or direct-cDNA experiments. Moreover, RT-PCR experiments failed to detect ~5000 genes that were detected by both PCR free approaches. Additionally, direct-RNA and direct-cDNA both contributes to the detection of ~1000 genes that are not detected in other experiments.
Those observations show that direct-RNA and direct-cDNA sequencing kits perform similarly in term of number of genes found and suggests that retro-transcription of the mRNAs does not affect gene ratios in cDNA libraries.
The lower number of genes found in RT-PCR experiments might be the result of sequencing a SL1-library instead of a poly(A) library, or it can be the result of amplification bias that will promote the detection of some genes and reduce it for others.

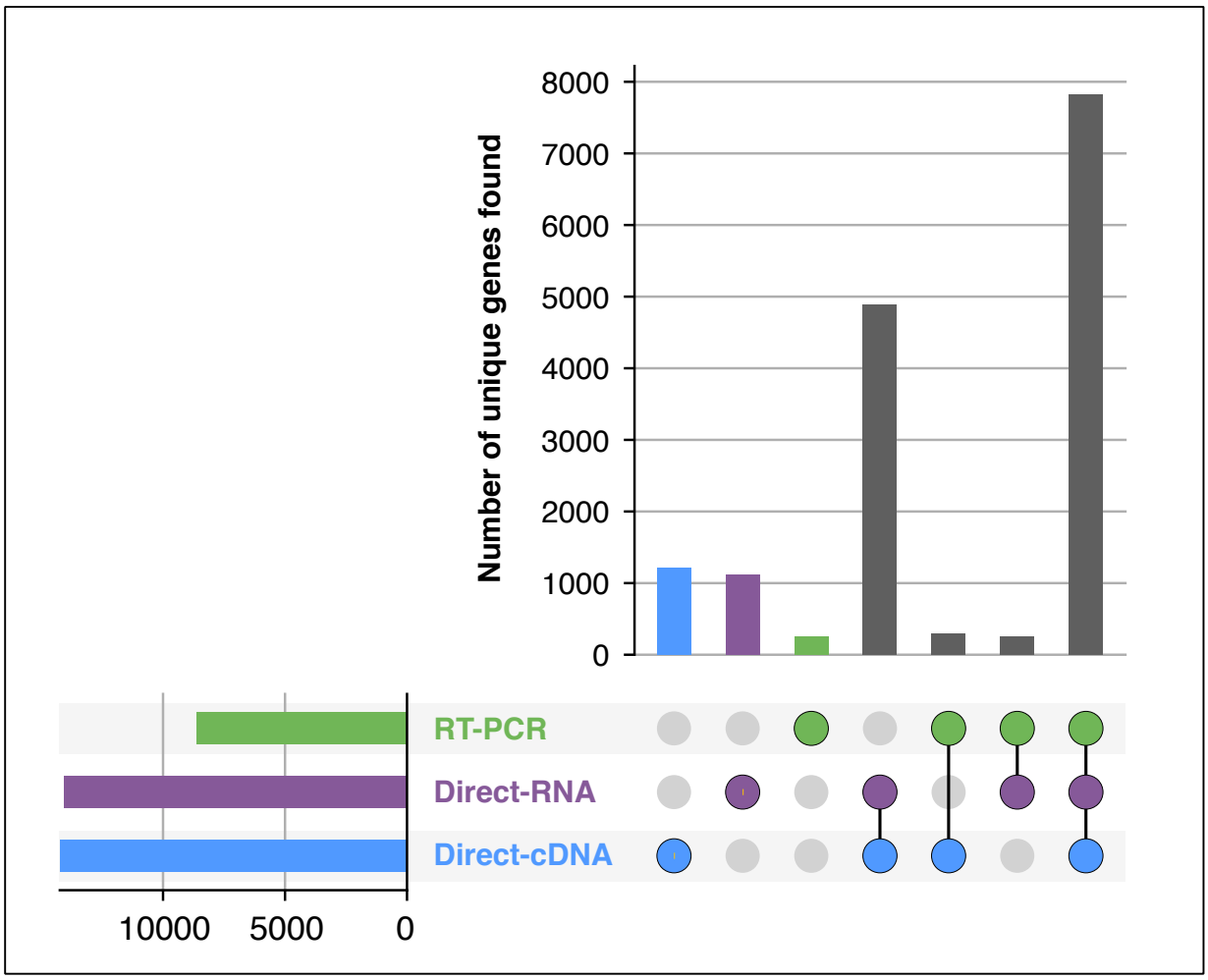


Figure 28 - Comparison of the number of genes detected between different type of sequencing experiments. A) Sequencing statistics per kit tested. The data from the duplicate experiments are pooled together. **B)** Upset plot for visualizing genes sets intersections between the different type of sequencing experiments.

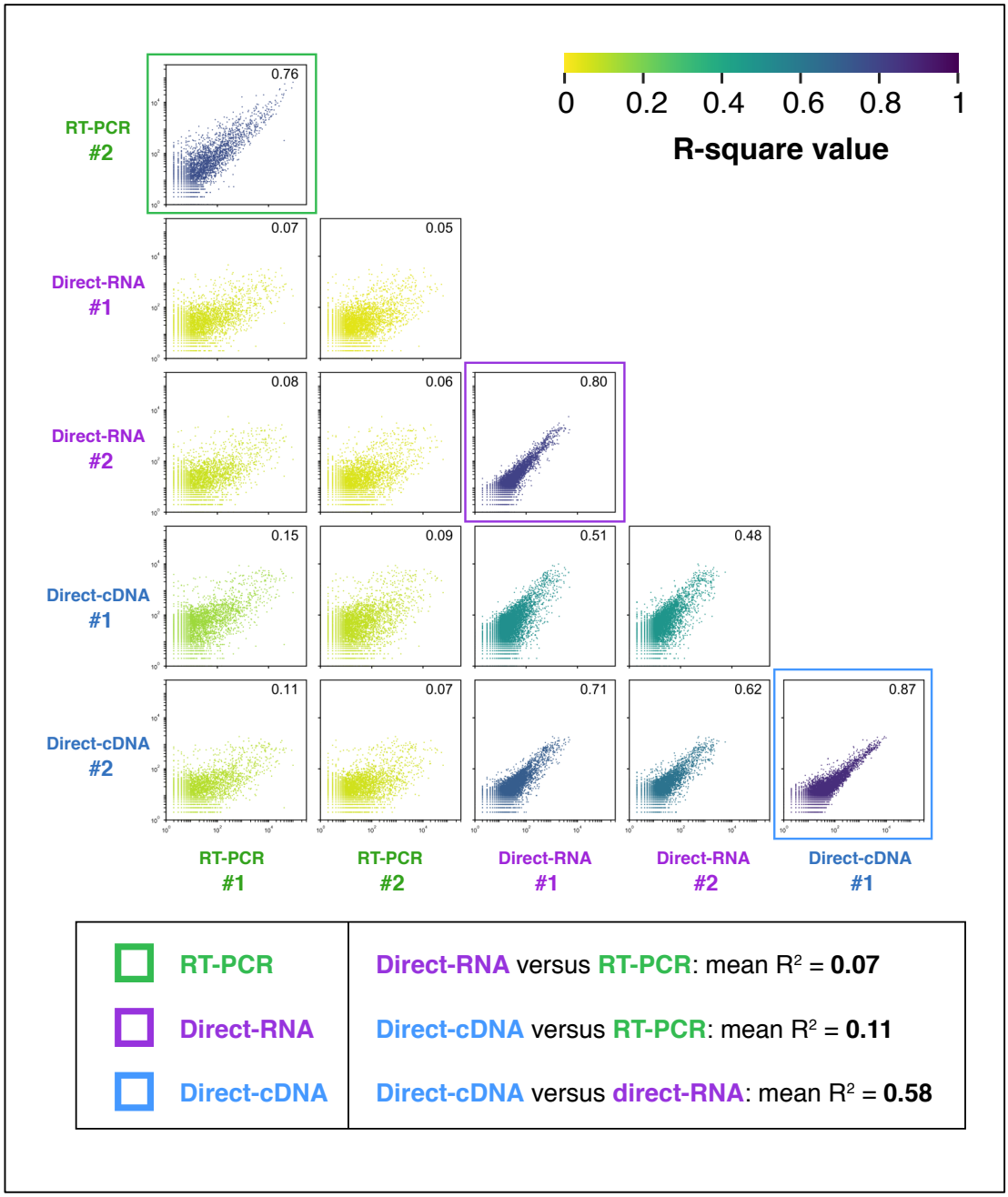


Figure 29 - Comparison of gene expression between each sequencing kit. Each plot is colour coded based on the R^2 value of the linear regression analysis.

Finally, we also looked at reproducibility between the different kits. To this end, the number of reads per genes found in each dataset was plotted in function of the number of reads per genes found in another dataset (**Figure 29**). Subsequent linear regression analysis was then performed to determine the similarity between them. The R^2 value of each comparison is used for color-coding of the different plots (yellow-green = poor correlation, blue-purple = good correlation).

From those results, we can see both RT-PCR datasets are very divergent from either the direct-RNA or direct-cDNA datasets, with R^2 values of about 0.11. However, the comparisons

between direct-RNA and direct-cDNA datasets showed a better correlation with an average R^2 value of 0.58. This results further confirm our previous observations regarding the similarity of direct-RNA and direct-cDNA datasets and seems to confirm the presence of altered genes ratios in PCR-based sequencing experiments.

f) Evaluation of RT and PCR bias:

The measure of potential bias introduced during library preparation was performed by counting the percentage of reads attributed to each gene. For each experiment, the percentage of total reads for a given gene was then plotted from the most expressed gene (bigger percentage of reads) to the less expressed gene (left panel). A second plot, representing the cumulative percentage of reads was also produced (right panel).

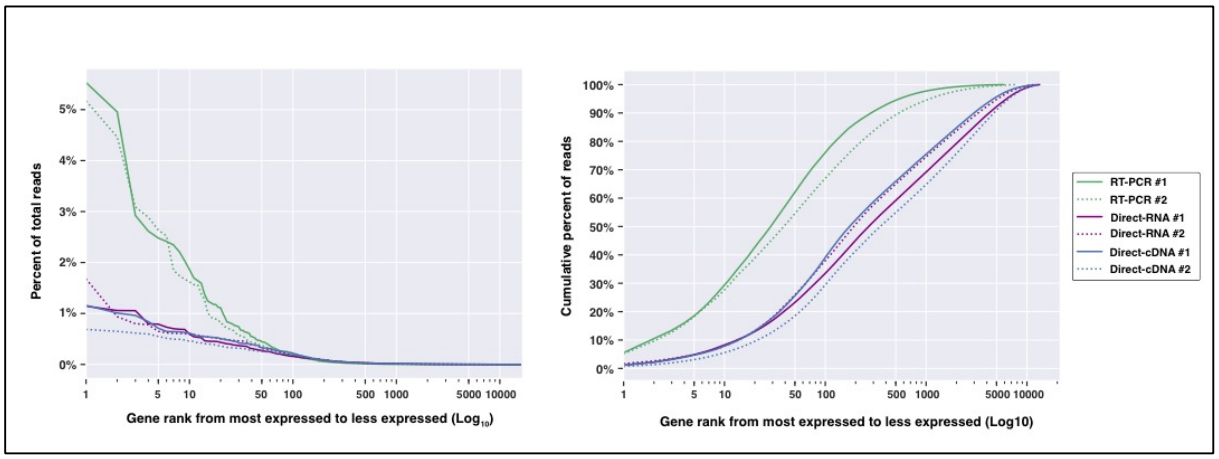


Figure 30 - Percentage of reads represented by top-ranking genes. The left panel show the percentage of reads attributed to each gene (from most to less expressed). The right panel show the total number of reads attributed to N top-ranking genes.

From these measures, we can observe top-ranking genes in RT-PCR experiments represent a bigger proportion of sequencing reads than top-ranking genes in direct-RNA and direct-cDNA experiments. In RT-PCR experiments, the 10 most expressed genes contribute to 30% of the total reads while it only represents ~10% of the reads in the other experiments.

Additionally, more than 90% of the reads comes from the 500 most expressed genes in RT-PCR libraries, while 90% of the reads in the other datasets are sufficient to capture up to 5 000 different genes. These observations highlight the introduction of PCR bias during library preparation.
Furthermore, the similarity of the curves between direct-RNA and direct-cDNA in **Figure 30**, combined with the scatter plot for gene expression (**Figure 29**), allow us to conclude that the retro-transcription step during direct-cDNA library preparation does not seem to influence gene expression ratios and provides an accurate measurement. Indeed, most commonly detected genes

(high gene count) are found more consistently between two experiences compared to lower expressed gene that still exhibit significant differences in term of observed gene count.

Nevertheless, even in the absence of PCR bias, we find that a significant proportion of reads originates from a small set of gene with ~50% of the read produced by the 500 most expressed genes.

g) Conclusion:

Considering the previous observations, it was decided to perform the transcriptome-wide analysis using the direct-cDNA sequencing kit.

The different comparisons that we performed to choose between the three kits indicate that direct-cDNA sequencing is best suited for performing RNA sequencing experiments in the nematode. DNA-based sequencing kits exhibit a higher throughput than RNA-based kits as well as a better quality of reads. Yet, the absence of PCR amplification with the direct-cDNA approach prevents the introduction of PCR bias that we have observed in the RT-PCR datasets.
Additionally, when testing the different kits, a study focused on the use of direct-RNA sequencing in yeast reported that reads generated with that kit were showing shorter 5' extremities compared to reads generated with direct-cDNA sequencing. With the objective of studying the presence of splice leader sequences on the 5' extremity of *C. elegans* mRNAs, we considered this as an important flaw for the approach and therefore decided to push forward with direct-cDNA sequencing.

2.1.2 - Splice leader sequences generate sequencing artefacts

Our three first direct-cDNA sequencing experiments were performed according to the supplier's recommendations on poly(A) purified RNAs using the provided SSP primer for 2nd strand synthesis. After mapping our reads onto *C. elegans* genome, we controlled their correct alignment by looking at them in Integrative Genome Viewer (IGV), and noticed an atypical behaviour not previously observed either without RT-PCT 1D or our direct RNA libraries.

A screen capture of direct-cDNA reads being aligned to a region of the genome is shown in **Figure 31.A**. We noticed large unaligned region (called soft-clipped regions) in 5' position on the vast majority alignment, relative to the gene orientation.
Additionally, we see a good coverage of the 3' region of the gene - as indicated by the presence of unaligned poly(A) regions at the end of the alignments - but a poor coverage of the 5' region. This is explained by the fact that we could observe a strong strand bias in favour of the antisense strand of the double-stranded cDNA molecule (purple reads in IGV) and indicates that we sequence our RNAs in a 3' to 5' fashion.

A model of a typical read obtained in this library is shown in **Figure 31.B**. The aligned region of the read is depicted in purple. On the 3' extremity there is a small soft-clipped region which corresponds to the first bases of the poly(A) tail as well as the sequencing adapters. On the 5' extremity, where we expect to observe the splice leaders and sequencing adapters, we see instead a much longer soft-clipped sequence.

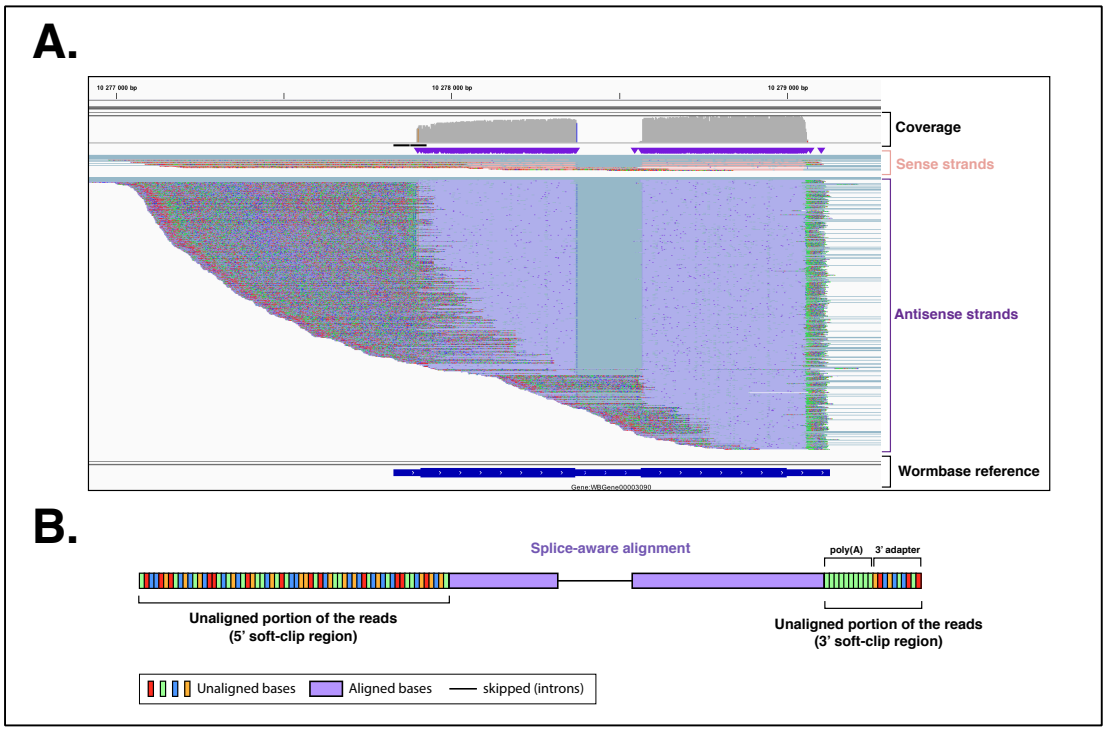


Figure 31 - *C. elegans* direct cDNA reads have long soft-clip regions and a strong strand bias. A) Example of reads aligned to a region of the genome. Purple alignments represent the antisense strand and pink alignments represent the sense strand of the cDNA molecule. **B)** Schematic view of the obtained reads. Unaligned regions (soft-clip) are present at both extremities. The 3' region is short and contains the poly(A) sequence (in green) and the 3' sequencing adapter. The 5' region is larger than expected.

We measured the number of reads coming from each cDNA strand by performing transcriptome mapping onto known reference RNAs sequences. Since the sequences are arranged in the 5' to 3' direction, we can determine the read's origin by looking at mapping statistics: if the read had to be reversed prior to mapping, it means the read was sequenced in the 3' to 5' direction and therefore comes from the antisense strand. We used an in-house python script along with the pysam library to loop over every transcriptomic alignment and determine if the read had been reversed or not. Out of 1,337,404 reads mapping on *C. elegans* transcriptome, we measured that 36,621 reads (2.74%) were coming from the sense strand of the ds-cDNA molecules and 1,300,783 (97.26%) from the antisense strand.

a) Large 5' soft-clipped regions in direct-cDNA read originates from antisense strands

To understand the origin of the long 5' soft-clip region (5SC), we measured the length of 5' extremities in every transcriptomic alignment and then plotted a histogram to look at size distribution. The same process was repeated for 3' soft-clip regions (3SC), which allowed us to compare both extremities. From the histogram, we can see that 5SC can range up to more than 1500bp. The mean size is about 540bp, while the median size is about 450bp. On the contrary, 3SC regions are much more consistent and range between 50bp and 70b, with both a mean and median size of about 60bp. This observed length is consistent with the size of the adapter (about 100bp long for direct-cDNA libraries) and is another indication that correct adapters are present at the 3' extremity but not the 5' extremity.

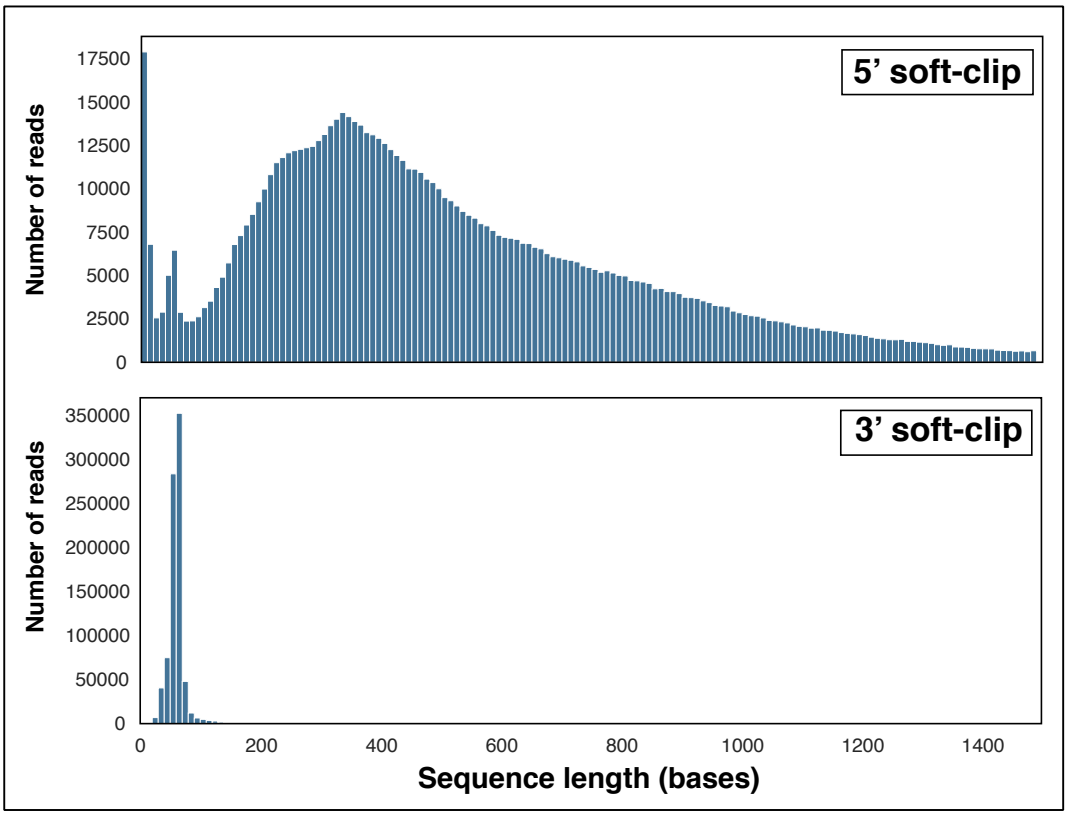


Figure 32 - Direct-cDNA reads only have long 5' soft-clip in *C. elegans* libraries. Top panel represent size distribution for 5' soft-clip sequence and bottom panel for 3' soft-clip sequences.

Since both strands of the cDNA should be sequenced in equal ratio under normal condition, we hypothesized that the 5' extremity of the cDNA molecule was no longer free for adapter ligation, resulting in the addition of a single adapter, on the 3' extremity.

By looking at 5SC length distribution in **Figure 32**, we also observed some reads with soft-clips sequence of about 60bp, just as in 3' extremities. Hence, we wondered if such soft-clip length could be the indication of reads exhibiting adapter sequences on their 5' extremity. To determine this, we plotted the size distribution of SC5 for reads originating from the cDNA sense strand and reads originating from the antisense strand (**Figure 33**). As a control, we also performed the same analysis onto a yeast dataset.

The result show that all the reads originating from the cDNA sense strand have a short soft-clip, just as in 3' extremities, meaning we were able to obtain those reads thanks to the correct ligation of our sequencing adapters on the 5' extremity. As seen previously, reads from the antisense strand mainly show a large soft-clip, yet we can still detect a certain number of reads with a short soft-clip as well. Compared with the average soft-clip size in sense reads, their size is slightly smaller, but this artefact is also seen in yeast and might be a side effect of sequencing the molecule from the 5' extremity and not the 3' extremity. However, by looking at both peaks, we can determine that there is about the same number of reads with a short soft-clip that originates from either strand. This indicates that cDNA molecules who had 5' sequencing adapter got both strand equally sequenced, as expected under such conditions, and is another proof that long soft-clip is a result of sequencing adapters lacking from the 5' extremity.

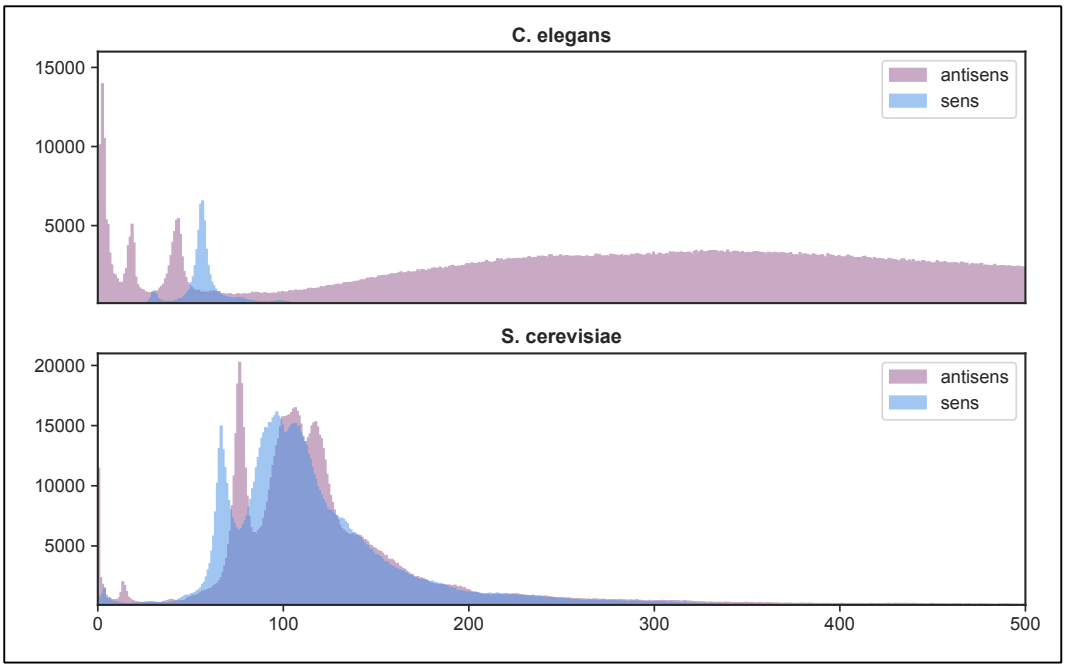


Figure 33 - Long 5' soft-clip originates from antisense strand in *C. elegans* libraries. Size distribution for 5' soft-clip coming from sense reads (blue) or antisense reads (purple). Top panel shows a *C. elegans* dataset and bottom panel a *S. cerevisiae* dataset.

b) 5' Soft-clip regions sequence corresponds to the sequence of the opposite strand

After confirming that 5' soft-clip sequences were coming from the antisense strand, we tried to determine if this sequence could correspond to a predetermined sequence. We noticed some reads had produced supplementary alignments. This happens when different regions of a single read can produce different alignments. The best alignment is classified as the primary alignment, and others alignments gets classified as supplementary alignments.

I used a python script to parse all of the alignments generated by our sequencing experiments and counted the number of reads that had zero, one or more than one alignment.
I established that ~70% of the reads had no supplementary alignments, ~30% of them had only one, while ~0.30% showed two or more.

We then isolated reads that produced a single supplementary alignment and extracted their mapping statistics. We could observe that ~96% of those reads have a supplementary alignment that corresponds to the same gene as their primary alignment, but in the opposite direction. Less than 1% of those reads matched the same gene in the same direction, and ~3% matched a different gene. Considering this observation, this indicates that the 5' soft-clip actually corresponds to the sense strand of the cDNA molecules: antisense strand gets sequenced first but, once it is finished, the sense strand gets sequenced as well, leading to concatenation of both strands sequence within a single read.

c) 5' Soft-clip regions sequence shows higher error frequency

Since it was not possible to detect supplementary alignments for all the reads, we wondered if this could be the result of poor sequence quality in the soft-clipped region, resulting in partial mapping in a small minority of reads. To do so, we investigated the base quality of the reads and compared the quality of bases in the 5' soft-clip and in the primary alignment (**Figure 34**).

When plotting the base-quality of a single read, we observed the quality was highly variable along the sequence. This behaviour might be inherent to nanopore technology which tends to produce very noisy reads. Hence, to potentially mitigate this effect, we plotted the base-quality of 100 reads mapping to the same transcript and then computed the mean Qscore at every position. The resulting curve was finally smoothed using the Savitzky-Golay algorithm implemented in the scipy package (Savitzky and Golay 1964; Virtanen et al. 2020). This algorithm is a digital filter that uses convolution to increase the precision of the data without distorting the signal tendency. It is commonly used in several studies, from the field of physics to analytical chemistry (Maddams and Mead 1982; We used the following parameters: window size = 31; polynomial order = 3.

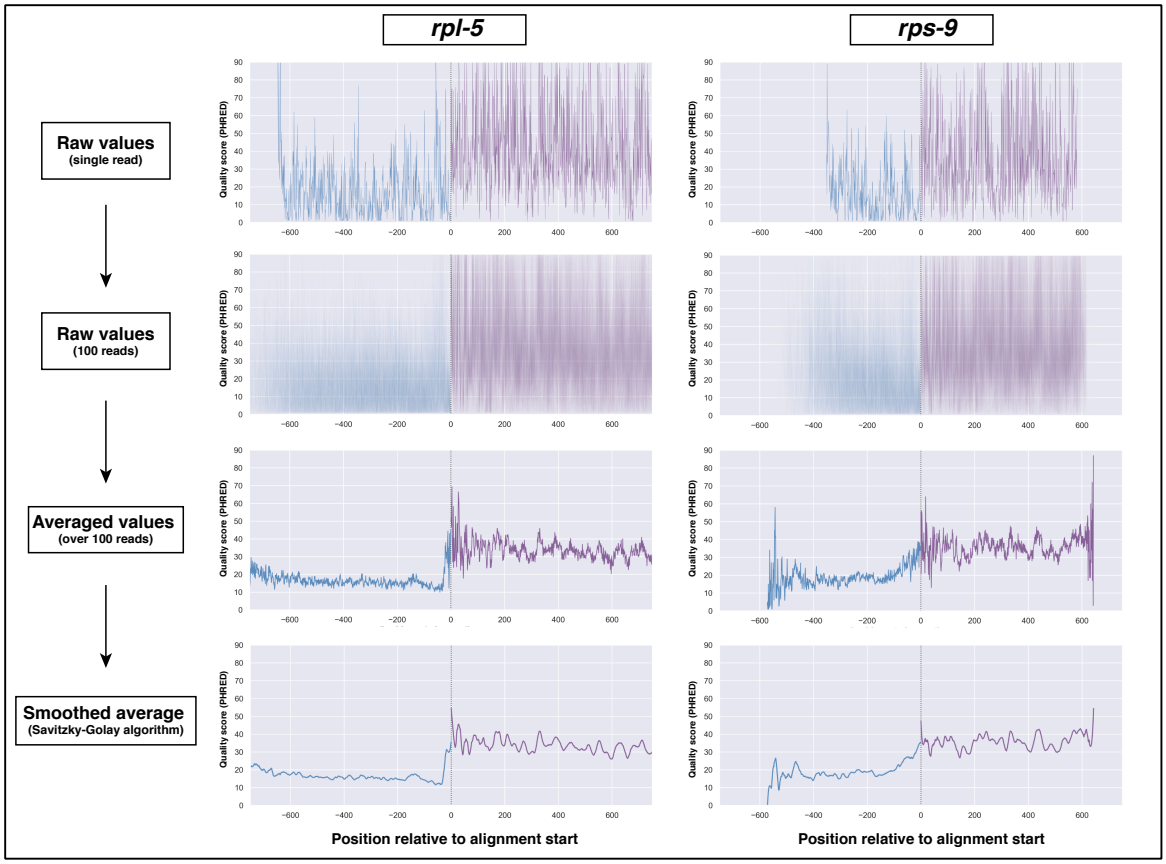


Figure 34 - Base quality in 5' soft-clip and primary alignment. Base quality for two genes highly expressed was investigated (*rpl-5* and *rps-9*). We measured the average base quality value over 100 reads and used a Savitzky-Golay algorithm for smoothing the data.

We find a significant difference in base quality between the 5' soft-clip regions and the aligned region of the read. 5SC regions have a mean Qscore of about 20 or less, which represents an error rate of about 1% or more. On the contrary, aligned regions show a mean Qscore of 30 or more, which corresponds to an error rate of less than 0.1% per base.
This analysis indicates that long 5SC likely consist of the missing sense strand reads with a lower quality as their cognate antisense strand.

This behaviour reminded us a discontinued kit from ONT called "2D library", where they used a hairpin sequence to physically link the two strands of the cDNAs molecules in order to sequence both strands at once. However, their latest motor protein was not able to properly read the second strand, so they stopped commercialization of this kit.

d) Splice Leader sequences mediate hairpin formation in direct-cDNA libraries

For 2D reads to occur, both strands need to be attached via a hairpin structure, and since we did not observe the same behaviour with yeast cDNAs, we suggested that the characteristic SL sequences of *C. elegans* can cause the formation of a hairpin during library preparation (**Figure 35.A**). To investigate this, I isolated the part of the sequence that is trans-spliced onto the RNAs and performed secondary structure predictions using the RNAfold web server. This analysis confirmed the ability of the SL1 sequence to form a hairpin, as well as some SL2 variants (**Figure 35.B**).

Since trans-splicing occurs at the 5' end of the mRNAs, I extracted reads with a SL1 sequence just before the beginning of the alignment, at the end of the 5SC. Knowing that nanopore reads - and more particularly 5' soft-clip regions - are very noisy, we preferentially isolated reads that exhibited a complete, error-free, SL1 sequence. After finding such reads, I extracted the region around the SL sequence, which span the last 100 bases of the 5SC until the first few bases of the primary alignment. They were then processed with RNAbow, in order to determine potential pairing within the sequence. As seen in **Figure 35.C**, we could confirm the detection of a SL1-mediated hairpin sequence within our reads. The sequence alignments indicate that the first two bases (GG) of the SL1 sequence pair up with two of the three C in position 12-14 (CCC) and that this pairing is sufficient to prime the second strand synthesis, thus forming a complete cDNA hairpin.

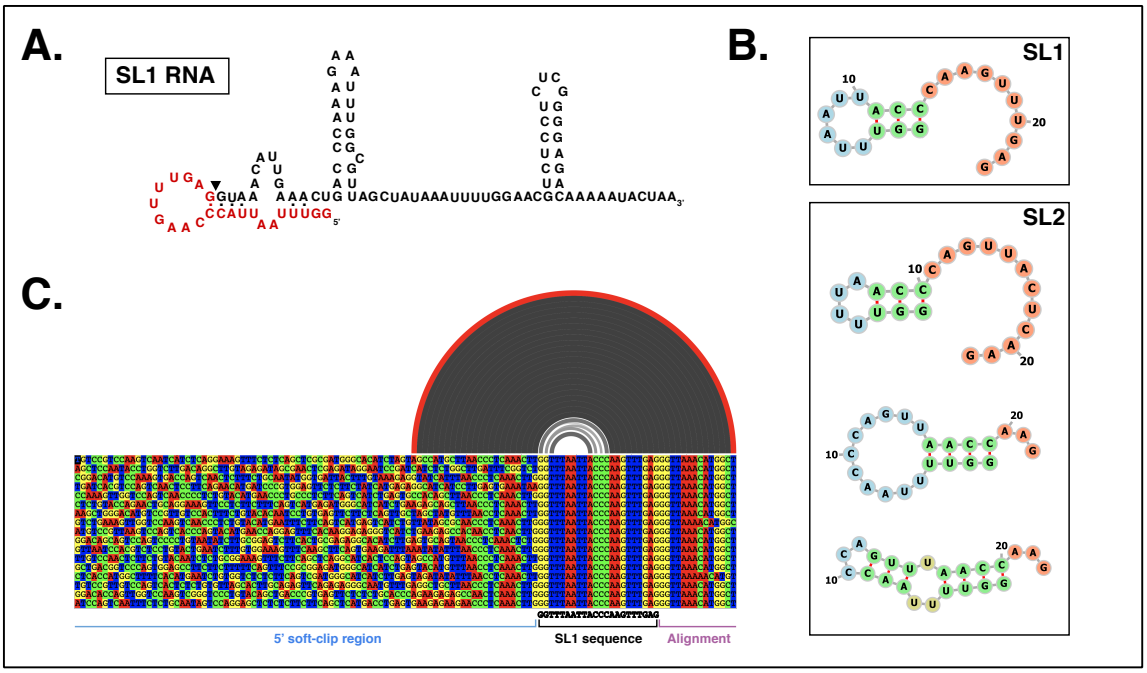


Figure 35 - Identification of an SL1 hairpin. A) SL1 precursor ncRNA. The sequence in red indicates the SL1 sequence found on mRNAs. Black arrowhead indicates the splice donor site. **B)** predicted RNA secondary structures for different splice leader variant sequences. **C)** Direct-cDNA reads presenting an SL1 hairpin.

Given these new results, we decided to test if it was possible to prevent hairpin formation during the preparation of the library. To do so, we generated new libraries by replacing the SSP primer (used in ONT's protocol for performing strand switching and 2nd strand synthesis) by either a SL1 primer or no primer at all. The rest of the protocol was carried out exactly as before.

As expected from a self-forming hairpin, not using any primer for 2nd strand synthesis, did not hinder our ability to sequence the resulting library as it produced reads just as in the other sequencing experiment. Furthermore, ONT indicates their adapters cannot be attached to single stranded cDNA molecules, confirming that we did generate double-strand cDNA molecules even in the absence of a leading primer for 2nd strand synthesis.

The reads were then basecalled and mapped onto *C. elegans* transcriptome and we measured again the proportion of reads originating from either strand. As expected, not using any primer further increased the strand bias initially observed in libraries generated with the SSP primer. On the other hand, the use of the SL1 primer made it possible to significantly reduce strand bias without completely preventing it. This result however confirms that it is possible to reduce strand bias by affecting the formation of the SL1 hairpin through the hybridization of a complementary sequence.

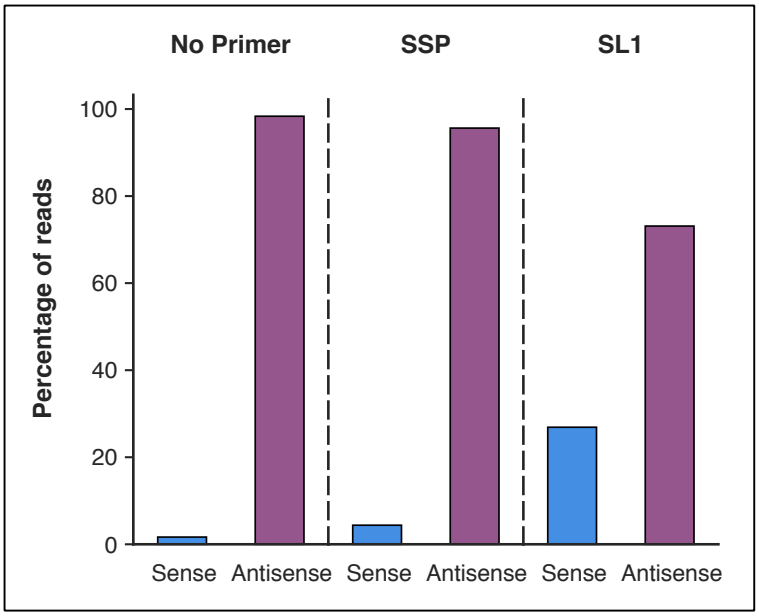


Figure 36 - Strand bias can be reduced by using a SL1 primer during library preparation. Strand bias was measured in dataset generated by using different primers for 2nd strand synthesis: SSP primer, SL1 primer or no primer.

e) Model: SL sequences creates artefact in direct-cDNA libraries

We propose the following model to explain how trans-spliced sequences at the 5' end of *C. elegans* mRNAs lead to the generation of atypical reads during direct-cDNA library preparation (**Figure 37**).

Splice leader sequence naturally forms hairpin as a secondary structure when single-stranded. However, after first strand cDNA synthesis, the structure is first disrupted by the RNA polymerase complex and then prevented by the presence of a complementary sequence. However, upon degradation of the RNA template, the complementary SL sequence is able to reform a new hairpin. The hairpin prevents the hybridization of the SSP primer for performing strand-switching, yet serves as a primer onto which the DNA polymerase can bind to. The 5' end being no longer free, sequencing adapters are only ligated onto the 3' extremity of the cDNA, which results in the sequencing of the antisense strand first.
However, when replacing the SSP primer with a SL1 primer, the hairpin structure gets disrupted and the 5' end is then available for adapter ligation.

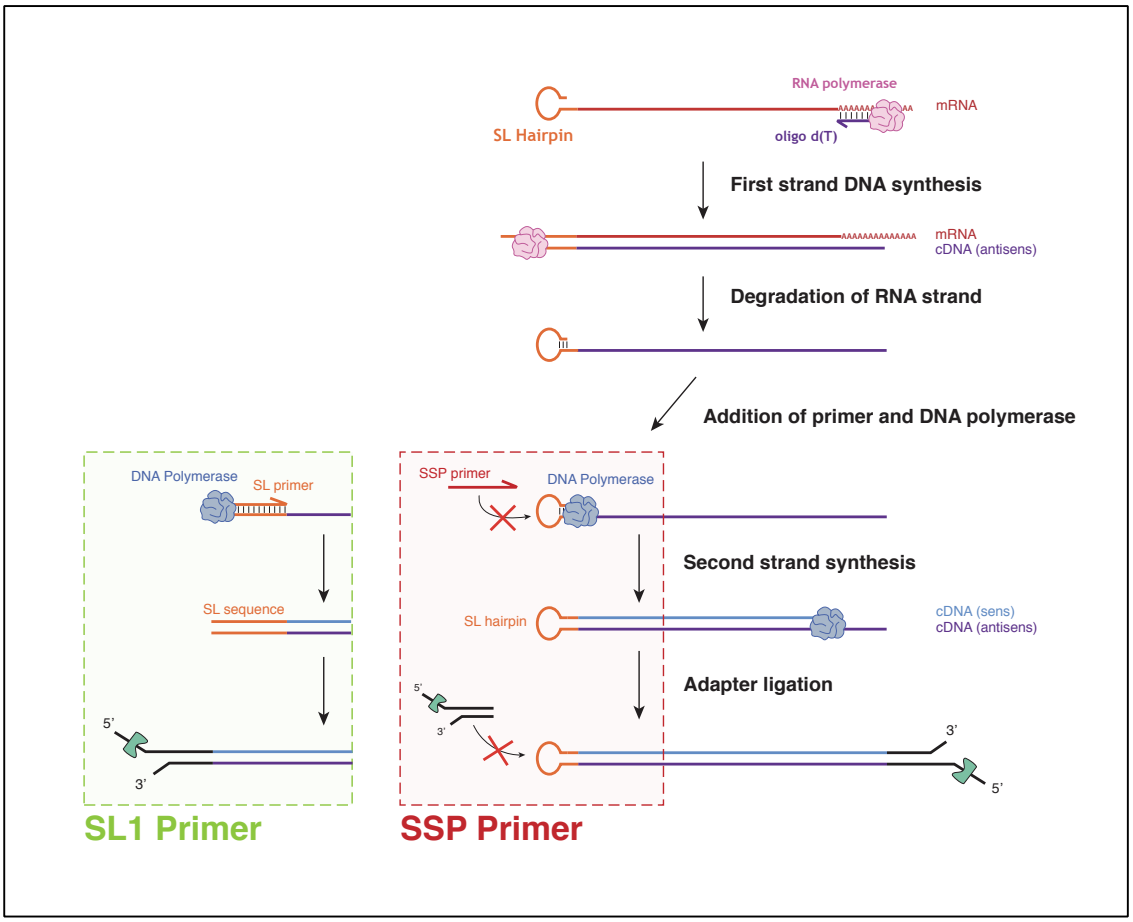


Figure 37 - Model for direct-cDNA library preparation of trans-spliced RNAs. During library preparation, the SL1 hairpin at the 5' extremity prevents the addition of the 5' adapter sequence. This phenomenon is attenuated when using the SL1 primer for 2nd strand synthesis.

Following sequencing of such atypical reads, the antisense cDNA strands on which is attached the unique motor protein is always sent first into the nanopore. Thanks to the helicase activity of the motor protein, both strands gets separated and sequencing initially occurs as expected. However, after sequencing of the hairpin region, the opposite strand is then pulled inside the nanopore as well. When the first bases exit the nanopore, the hairpin gets reformed and both strands start to pair again (**Figure 38**).

It has been confirmed by Oxford Nanopore that this behaviour significantly affects sequencing quality by applying constraints onto the rest of the molecule. As the rest of the sequence passes through the pore with a different speed, it is no longer possible to accurately basecall the sequence with the same parameters than normally used, explaining the poor base quality of this part of the read. ONT discontinued their 2D sequencing kit because it wasn't compatible with the current motor proteins they are using.

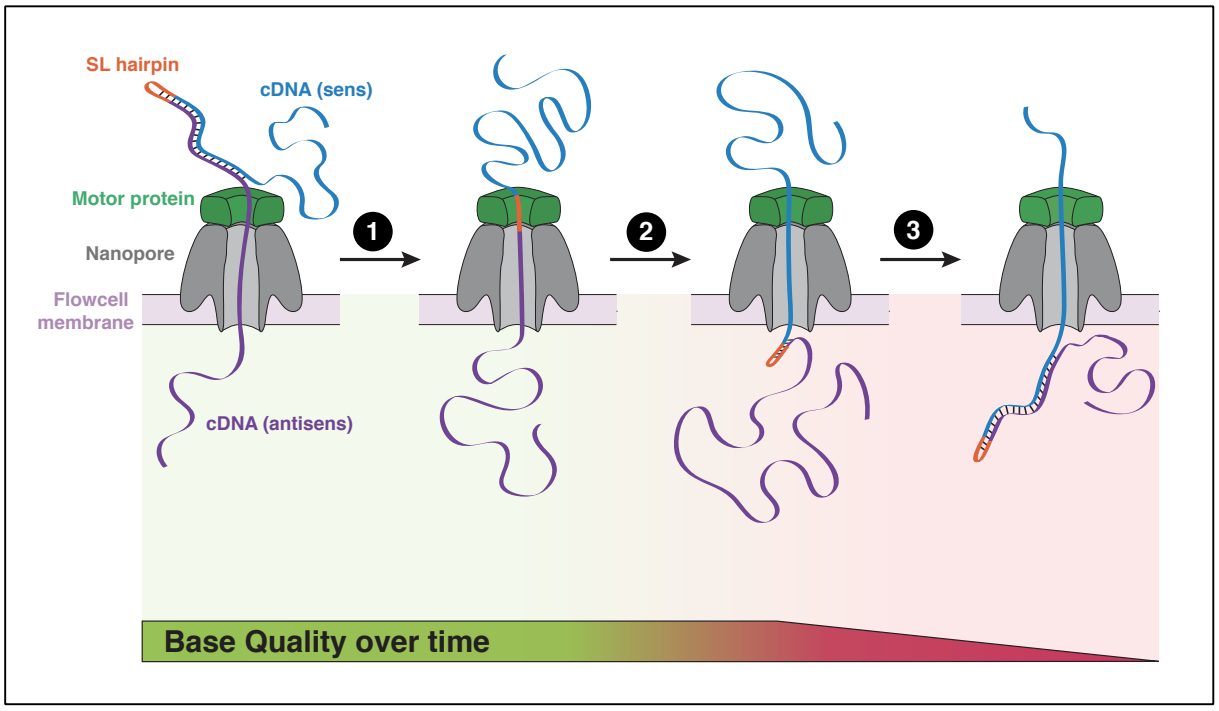


Figure 38 - Hairpin reads affects sequencing behaviour. The antisense strand is sequenced first. Then, the sense strand is pulled inside the pore by the hairpin linking both strands. As both strand gets paired again, physical constraints are applied onto the rest of the molecule, affecting sequencing speed and ultimately base quality.

As an additional control, we decided to attempt a transcriptome sequencing experiment on *Leishmania tarentolae*. This species is known for using trans-splicing as well and we could observe a similar strand bias: about 95% of reads are originating from the antisense strand.

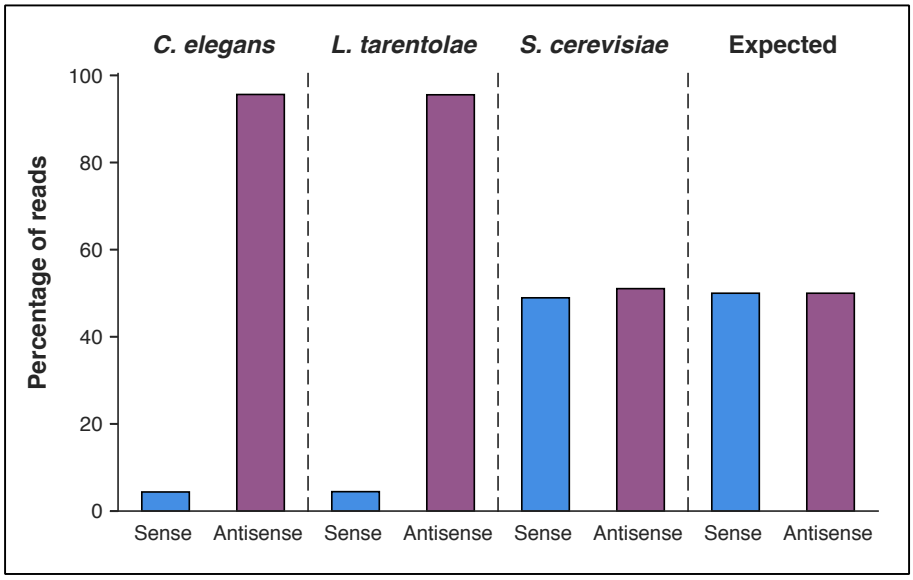


Figure 39 - Species with trans-splicing display strong strand bias with direct-cDNA sequencing.

Interestingly, the splice leader sequence in *L. tarentolae* is very different from the SL1 sequence found in *C. elegans*, however its predicted secondary structure shows an even stronger pairing between bases in positions 15-23 and 27-34 (**Figure 40**). However, we observed similarities with the SL2 variants found in *C. elegans*. This observation might indicate that all SL RNAs have retained the ability to form hairpins.

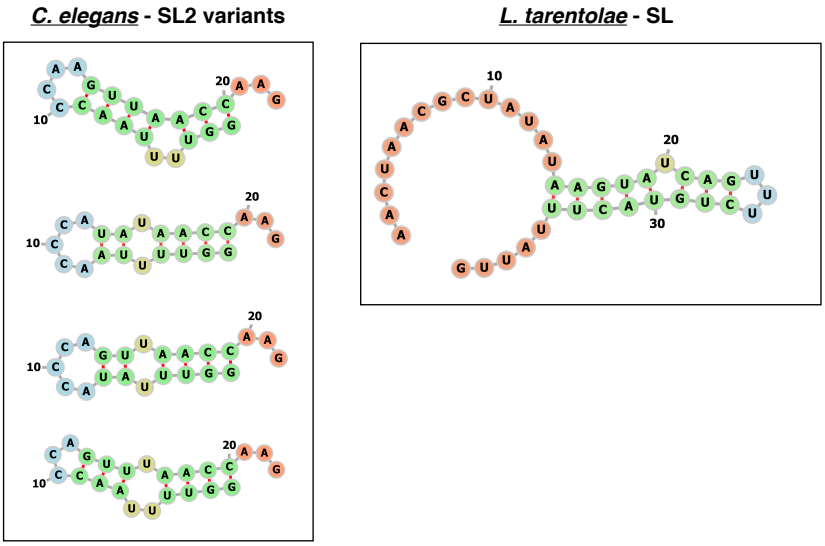


Figure 40 - Secondary structure prediction for *C. elegans* SL2 variants and *L. tarentolae* SL sequence.

2.1.3 - Description of the direct-cDNA datasets

a) Overview of the different datasets

During this project, a total of 12 direct-cDNAs sequencing experiments were performed. Those 12 experiments can be split into four different groups:

- SSP [polyA]: Three datasets generated by sequencing poly(A) RNAs. Second strand synthesis was performed using the universal SSP primer that is provided with ONT library kits. These include the original experiments performed to assess the performances of the different ONT kits in the conditions described by the manufacturer.

- SSP [sl1]: Three datasets generated by sequencing SL1 RNAs. Second strand synthesis was performed using the universal SSP primer.

We tested these two strategies as a potential way to find new genes for which trans-splicing had not been detected previously to pursue the observations from team that trans-splicing is more prevalent than previously thought and that genes for which we could not detect trans-splicing were low expressed genes (Tourasse et al, 2017). Furthermore, the isolation and sequencing of SL1 RNAs is directly linked to another project which consists in generating tissue-specific transcriptome maps (see **section 2.3**).

- SL1: A single dataset generated by sequencing poly(A) RNAs. In this experiment, second strand synthesis was performed using a SL1 primer (see **section 2.1.2** for the details of that experiment).

- NoPrimer: 5 datasets generated by sequencing poly(A) RNAs. In those experiments, we did not add any primer for Second Strand Synthesis (see **section 2.1.2**).

We generated these two additional datasets during the study of the sequencing artefacts observed in our reads.

For each experiment, detailed statistics regarding the number of reads obtained after basecalling, the number of reads mapping to the genome and to the transcriptome, or the number of reads found associated with a splice leader (SL) sequence is noted in **Table 3**.

Name	Template	2nd strand synthesis	#	Basecalled reads	Genomic alignments	transcriptomic alignments	SL reads found	Genes found	SL genes found
SSP [polyA]	poly(A) RNAs	SSP primer	1	1 067 062	926 986	856 699	219 712	12 562	7 667
			2	372 188	302 164	265 555	60 429	12 376	6 526
			3	428 451	269 505	215 150	48 967	11 537	6 209
			Total	1 867 701	1 498 655	1 337 404	329 108	14 662	9 535

Name	Template	2nd strand synthesis	#	Basecalled reads	Genomic alignments	transcriptomic alignments	SL reads found	Genes found	SL genes found
SSP [sl1]	SL1 RNAs	SSP primer	1	805 214	563 132	329 046	64 545	14 074	7 305
			2	203 384	122 662	75 874	12 020	9 511	3 267
			3	2 698 484	2 003 969	520 347	98 598	12 235	7 021
			Total	3 707 082	2 689 763	925 267	175 163	15 489	9 300

Name	Template	2nd strand synthesis	#	Basecalled reads	Genomic alignments	transcriptomic alignments	SL reads found	Genes found	SL genes found
SL1	poly(A) RNAs	SL1 primer	1	7 811 076	6 776 420	6 015 856	3 233 672	14 637	11 898

Name	Template	2nd strand synthesis	#	Basecalled reads	Genomic alignments	transcriptomic alignments	SL reads found	Genes found	SL genes found
NoPrimer	poly(A) RNAs	No Primer	1	330 272	149 490	106 081	25 766	6 832	2 835
			2	3 238 319	2 666 799	1 962 645	488 328	13 208	8 051
			3	630 506	514 711	398 165	96 084	9 910	4 997
			4	99 031	77 991	54 418	12 575	5 432	2 040
			5	575 203	456 843	353 154	95 618	9 768	5 249
			Total	4 873 331	3 865 834	2 874 463	718 371	13 961	8 949

Table 3 - Description of 12 direct-cDNA experiments.

b) Reproducibility between each dataset

For each sequencing experiments, we controlled their reproducibility with their duplicates (same group) and with the other groups by comparing gene's level of expression. For each comparison, we plotted gene's reads in one experiment versus the other. Similarity between the two sample was then measured by performing a linear regression analysis. The results are showed as a matrix of scatter plots in **Figure 41**. For quick visualization, each plot was color-coded depending on the R^2 value of the linear regression analysis. The R^2 value was also added in the top right corner of each plot.

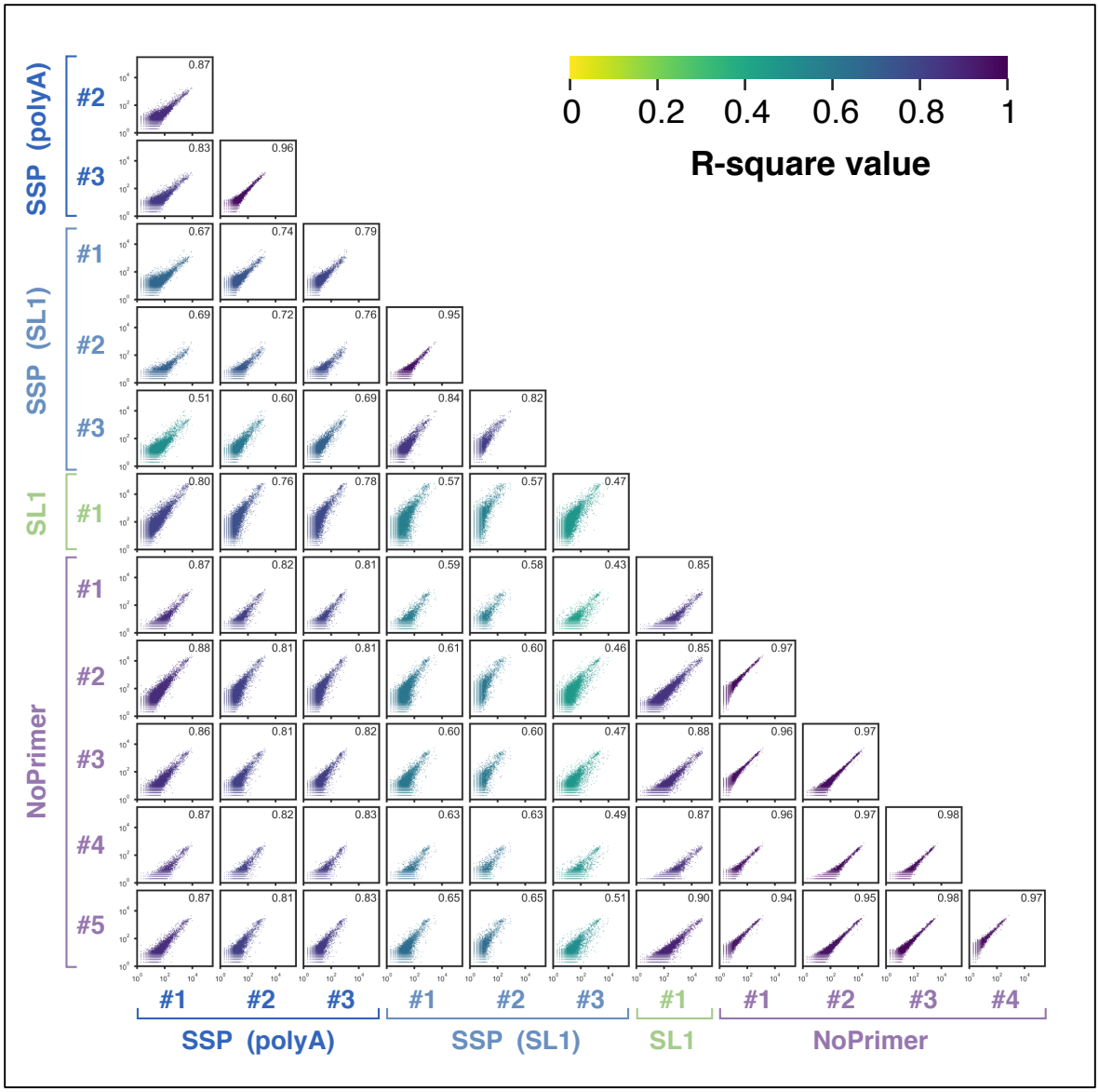


Figure 41 - Reproducibility between different direct-cDNA experiments. The colour of each plot represents the R^2 value of the linear regression analysis (the value is also noted in the top right corner).

As initially observed when we compared different sequencing experiments in **section 2.1.1**, duplicates direct-cDNA experiments produce similar results in term of gene expression level. Duplicates experiments in SSP [polyA] and SSP [sl1] datasets show R^2 values situated between 0.85 and 0.95, confirming a very good reproducibility. The NoPrimer datasets performed even better with a minimum R2 value of 0.94 and a maximum value of 0.98, suggesting very homogenous datasets.

Furthermore, poly(A)-based sequencing experiments tend to show a good correlation across the different sequencing protocols. When compared together, all three groups - SSP [polyA], SL1 and NoPrimer - show R2 values situated between 0.8 and 0.9. This suggest that swapping the SSP

primer with a SL1 primer, did not affected our detection level but only changed the characteristic of the reads (less SL1 hairpins observed). Incidentally, omitting the SSP primer for 2nd strand synthesis only increased the percentage of hairpin observed without affecting the number of genes detected.

Finally, as is expected from comparing experiments performed on different population of RNAs, the SSP [sl1] dataset present lower R^2 values when compared with any experiments of the other three groups (between 0.43 and 0.79). This observation confirms that different populations of RNAs were pulled down depending on using SL1 or oligodT primers as bait. It is unclear why this is the case since our results indicate that all mRNAs carry a SL. It is possible that this was caused by counter selection of SL2 genes.

c) Gene sets intersection for the different groups of direct-cDNA experiments

We then evaluated the differences between each dataset by looking at gene sets intersection between each of the groups (**Figure 42**). For this analysis, data from duplicates experiments were pooled together. The result was plotted as an Upset plot for easier visualization. From this plot, we can observe that a majority of the genes are detected in all four groups (~11 500 genes).
We also observe the SSP [sl1] dataset lead to the detection of ~1 600 genes that are not found in any of the other group despite its relatively low number of protein-coding reads (900K versus 1.3M, 6M and 2.9M for group SSP[polyA], SL1 and NoPrimer respectively).

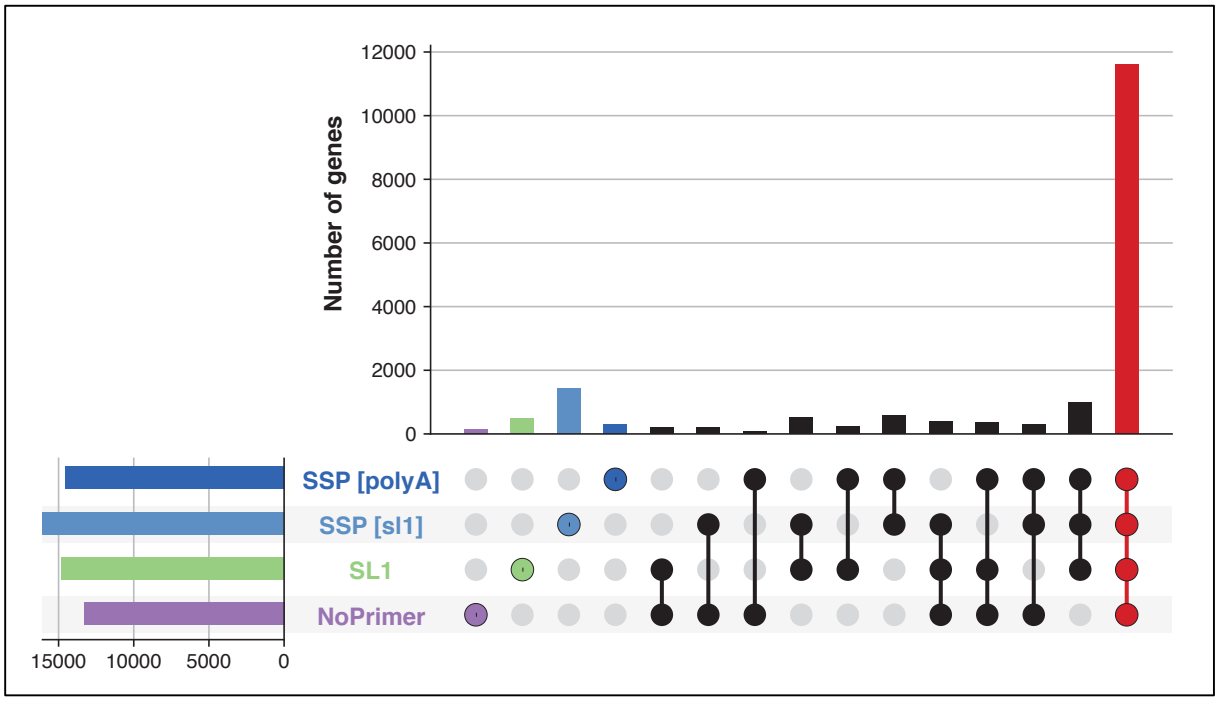


Figure 42 - Upset plot for visualization of gene sets intersections between the four types of experiments: SSP [polyA], SSP [sl1], SL1 and NoPrimer. Genes found in all four groups are highlighted in red, and genes uniquely found in one dataset are coloured based on their group.

d) Summary of sequencing statistics

Following the previous observations, we decided to pool together the data from the four different groups in order to study the trans-splicing status of the different genes detected. The rationale of this approach being the increase of gene coverage by the addition of several sequencing experiments combined with the fact that each group present genes that are not detected in any of the others.

The sequencing summary of the pool of all 12 direct-cDNA experiments is represented below. Statistics regarding the detection of SL reads and SL genes will be addressed in **section 2.2.1**.

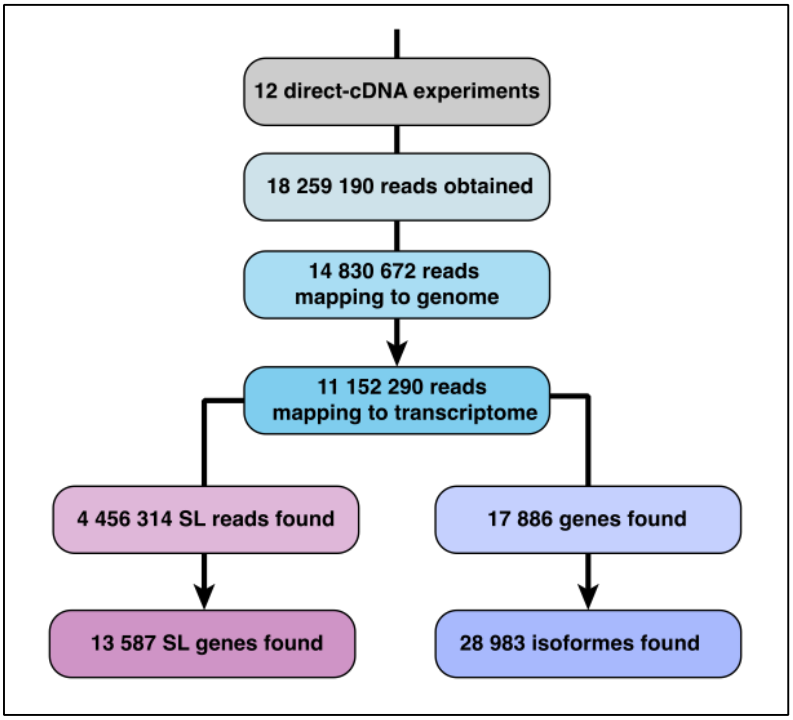


Figure 43 - General summary of all 12 direct-cDNA experiments generated during this project.

2.2 - Trans-splicing is a pervasive mechanism

As mentioned in previous sections, nanopore reads have a 10-15% error rate. Combined with the fact that 5SC sequences are noisier due to the presence of a hairpin, identifying a splice leader sequence of 22nt becomes a challenge. In this section is presented our attempts at finding splice leader sequences and then classifying the trans-splicing status of the different genes detected by direct-cDNA experiments.

In this section, the manipulation and exploration of the sequences, as well as the development of the different algorithms was performed using *ad hoc* python scripts that I wrote myself.

2.2.1 - Searching for splice leader sequences in nanopore reads

a) Direct search: finding error-free SL sequences

Our first approach was to count the number of reads aligned onto *C. elegans* transcriptome for which we could detect an error-free splice leader sequence. Since several splice leader sequences can be found in *C. elegans* mRNAs, we decided to search for the SL1 sequence as it is found in most of the messengers. Given that trans-splicing occurs on the 5' extremity of the mRNAs, the sequence is expected to be found just upstream the start of any transcriptomic alignment. Therefore, we extracted the last hundred bases before the alignment start (corresponding to the end of the 5' soft-clip region) We also included the two first bases of the aligned sequence since the donor site of SL1 and the acceptor site of the messenger are the same sequence (AG), meaning these two bases would be considered as part of the alignment (**Figure 44.A**). This sequence is referred as 5SC in the rest of this section.

Following extraction of this sequence, we performed an exact search for the 22 letters. We analysed the three sequencing experiments performed with the SSP primer on poly(A) mRNAs and found a match for 0.81% of all the reads in the first experiment, 1.19% in the second and 2.44% in the last (**Figure 44.B**).
This low incidence of perfect match is much lower that what would be expected form the methods error rate however we have determined that the presence of a hairpin affects base quality in the 5SC region. This result prompted us to develop another approach for searching SL1 sequences in our dataset.

To tackle this problem, we decided to look for shorter SL1 sequences. First, we chose to look for shorter SL1 sequences by removing the 3 first bases of the SL1 sequence and performing the same search as before. The operation was then repeated until we reached a minimal SL1 sequence of 7bp. The rationale behind this approach is that we can accept shorter sequences since the presence

of a SL is indicated by the long 5SC sequences. As seen in (**Figure 44.B**), this method allowed us to find more sequence, reaching a total of 15-20% of the reads in all datasets when searching for the last 7nt of the SL1 sequence.

We also repeated the same procedure on the dataset generated using the SL1 primer for initiation of the second strand synthesis. In this dataset, we found a significantly larger population of reads (18.1%) when searching for the full SL1 sequence. Furthermore, searching for shorter SL1 sequence allowed us to detect up to a maximum of 43.9% of all the reads when searching the last 7nt of SL1. Several factors can explain the observed differences:

- The presence of the hairpin in SSP libraries makes it harder to correctly sequence the 5' extremity of the cDNA. As mentioned before, 5' soft-clip regions are noisier than the aligned region due to modified physical constraints on the helicase that affects molecule speed through the pore and thus hinders our ability to easily detect SL sequences.

- Reads with no hairpin in SL1 library can be sequenced from their 5' extremity first (sense strand) allowing for a better coverage and a better quality of that region - unlike reads with hairpins who are sequenced from the poly(A) first - and increase our ability to detect short sequences.

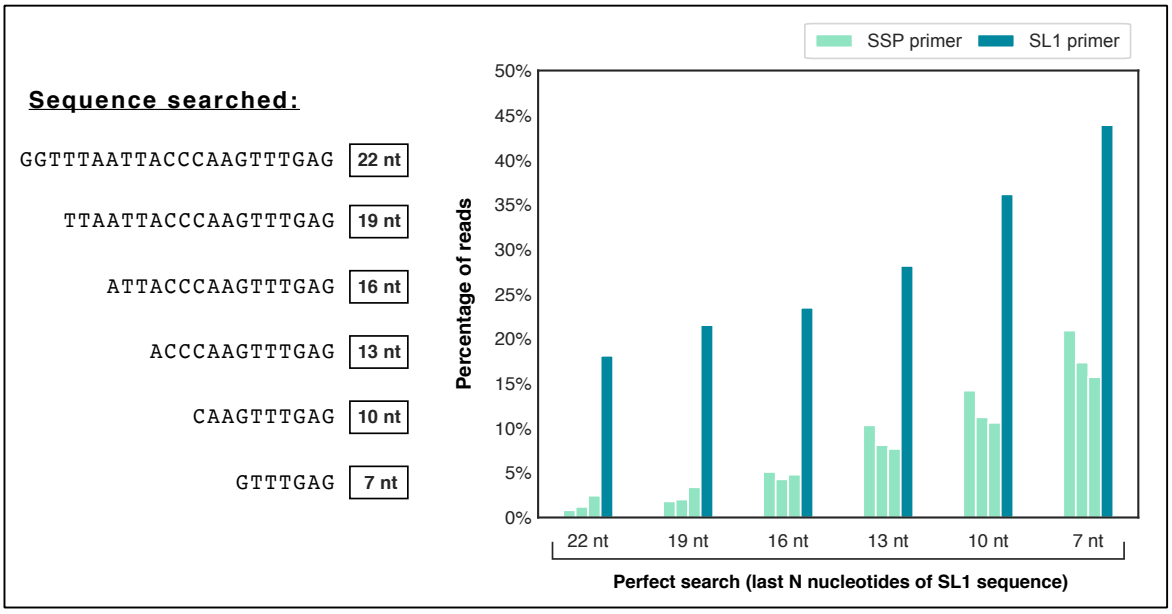


Figure 44 - Search of an unaltered SL1 sequence. A) Example of an SL sequence located at the very end of the 5' soft-clip alignment. **B)** Percentage of reads found with a perfect SL1 sequence of length N.

b) Imperfect SL sequences search

We sought out a method for finding SL sequences while allowing the presence of errors in the sequence. Nanopore technology being relatively recent, no software had been specifically developed for finding short sequences in noisy reads, therefore I decided to develop my own method. The method was inspired by a pre-existing python script called pychopper that was developed for removing adapter sequences from nanopore reads (available at: https://github.com/nanoporetech/pychopper).

As done in pychopper, I used the parasail python library in order to map a short sequence (the SL1 sequence) onto larger sequences (the 5' soft-clip end of each read). I used the semi-global alignment method of parasail, as it allows for finding the best partial alignment between two sequences. Best alignment and scoring of the alignment is made based on a substitution matrix. I used the following parameters:
match = 1 | mismatch = -1 | gap opening = -2 | gap extension = -1

An example of semi-global alignments of SL1 sequence against different reads, and their associated score, is given in **Figure 45**.

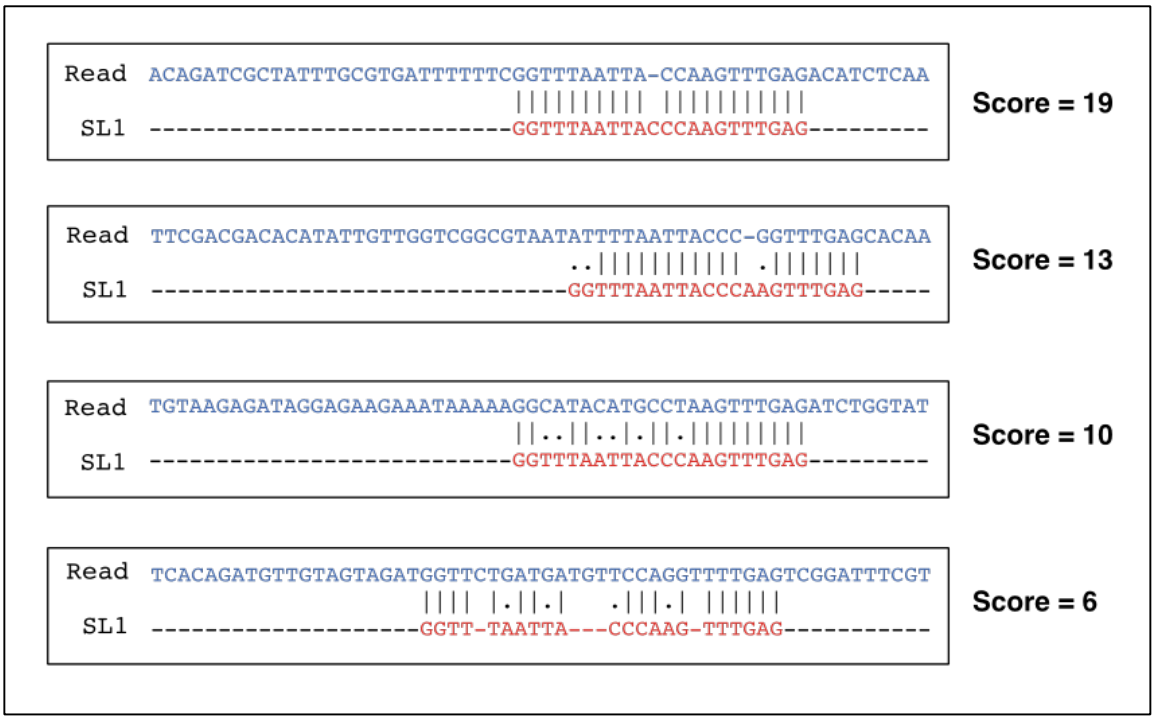


Figure 45 - Examples of semi-global alignment between SL1 sequence and 5SC regions. The end of the 5SC sequence is shown in blue and the SL1 sequence in red. Vertical bars between the two sequences represent a match between two bases, a dot represents a mismatch and a dash represent a gap in the sequence.

The method works as follows:

1) A sensitivity parameter (S) ranging between 0 and 1 is defined at the start. This parameter is used for calculating the minimum alignment score accepted based on the length (L) of the sequence searched:

$$L \times S = \ minimum\ accepted\ score$$

Since the maximum score of any given alignment is equal to its length (a sequence of n bases that is perfectly mapped will produce a score of n), by looking at alignments who generated a score close to the maximum possible score, we can confidently detect sequences that are similar to the sequence searched. With this method, we can easily search for all the SL sequences (SL1 and SL2 variants) and then retain only the one that generated the higher score.

By default, we used a sensitivity parameter of 0.7 (see **Figure 46**), in order to retain sequences that reached 70% of the maximum possible score. This score is termed "minimum accepted score".

2) If the alignment score is equal or superior to the minimum accepted score, the match is accepted, the score added to a list, and another SL sequence is evaluated.
If the alignment score is below the required score, the SL sequence is shortened and a new alignment is performed. The process is repeated until we reach a 7nt SL sequence. If no match is found, the evaluated sequence is considered absent from the read.

3) Once all of the SL sequences have been evaluated, the best scoring sequence is accepted. In case several sequences scored equally, two case scenarios can happen:
If only SL2 variants scored equally, we accept the match as a SL2 sequence without being able to identify precisely which variant.
In case the SL1 sequence scored as well as a SL2 sequence, we accept the match as an undetermined SL sequence.

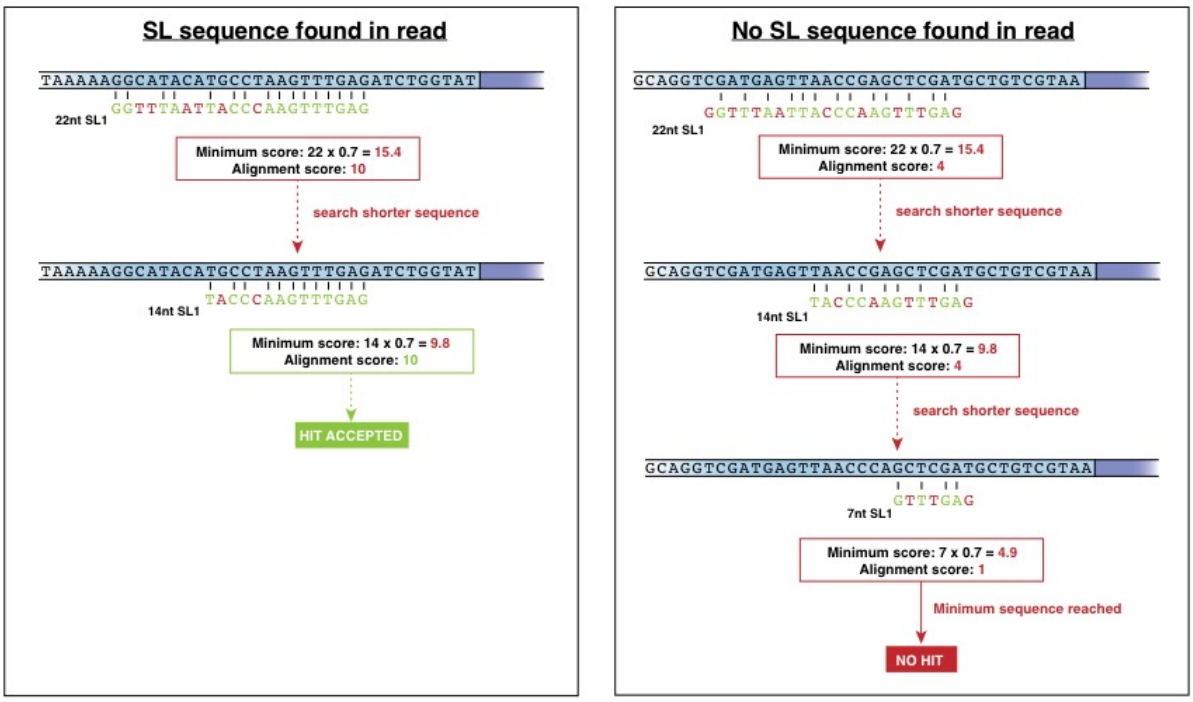


Figure 46 - SL search using our custom python algorithm. A search on a read containing a SL1 sequence is depicted on the left panel, and on a read with no SL1 sequence on the right panel. For each step, the minimum score is calculated based on the length of the sequence searched and compared with the alignment score of the sequence. If the minimum score is not reached the sequence is shorten and a new alignment performed.

In order to test the efficiency of our method, we randomly extracted a hundred alignments from two highly expressed genes (*rpl-5* and *rpl-41.2*) from two different libraries (SSP [polyA] and SL1 libraries). As done previously, we extracted the end of the 5' soft-clip sequence and the first two bases of the alignment and then measured the number of sequences for which we could detect a SL1 sequence using different approaches:

- **Manual count:** visual inspection of each sequence.
- **Perfect Search:** same protocol as described in **section 2.2.1.a**.
- **Blast:** The SL1 sequence was used to generate a local BLAST database and against which we blasted all sequences (parameters: short-blast, E-value < 0.005).
- **In-house SL search:** different sensitivity parameters were tested (0.6, 0.65, 0.7, 0.75 and 0.8).

The number of reads found with each approach (**Figure 47**) was then compared to the number of reads that could be detected by performing a visual inspection of the sequence.

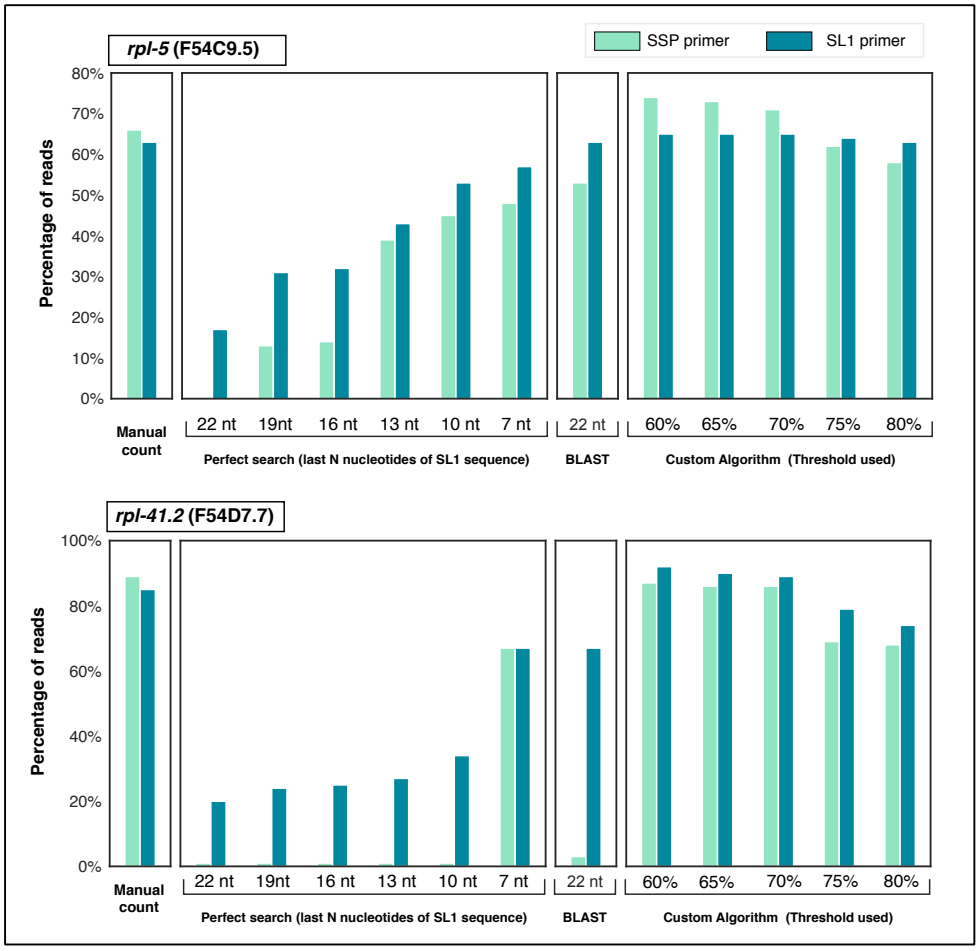


Figure 47 - SL search using four different methods: Manual count, Perfect search, BLAST and In-house SL search.

From the results, we can see our method allowed us to find more reads compared to the other approach. We also determined that a sensitivity parameter of 0.7 allowed us to detect almost the same number of reads than by performing visual inspection. Higher values were too stringent which resulted in fewer reads identified while lower values generated more hits than we could account for, suggesting false positives hits.

In order to prevent the insertion of false positive hits, we decided to also evaluate each hit based on its distance to the end of the sequence. As explained previously, SL sequences are expected to be found at the very end of the extracted sequence, however due to the nature of our reads (and in particular the presence of SL generated hairpins) the mapping of the extremity is prone to errors and part of the 5' extremity can be included in the 5' soft-clip sequence (**Figure 48.A**). Additionally, transcriptomic alignments are dependent on the quality of the current annotations available, some of which are the results of bioinformatics predictions, hence affecting the distance between the alignment start and the position of the SL sequence in the 5' soft-clip region.

By calculating the distance between a SL hit and the beginning of the aligned sequence (referred here as "distance to start"), we observed that strong SL matches (alignment score above 15) are almost exclusively located at the very end of the soft-clip region (distance to start < 20bp) whereas weaker matches (score less than 10) are found all along the sequence (**Figure 48.B and C**). This suggests that low scoring matches can correspond to random sequence similarity. Yet, we can observe a larger proportion of hits near the very end of the sequence, as expected from true positive hits. Therefore, we decided to exclusively accept low scoring hits (score < 10) only when detected near the end of the sequence (distance to start ≤ 5) and reject all others.

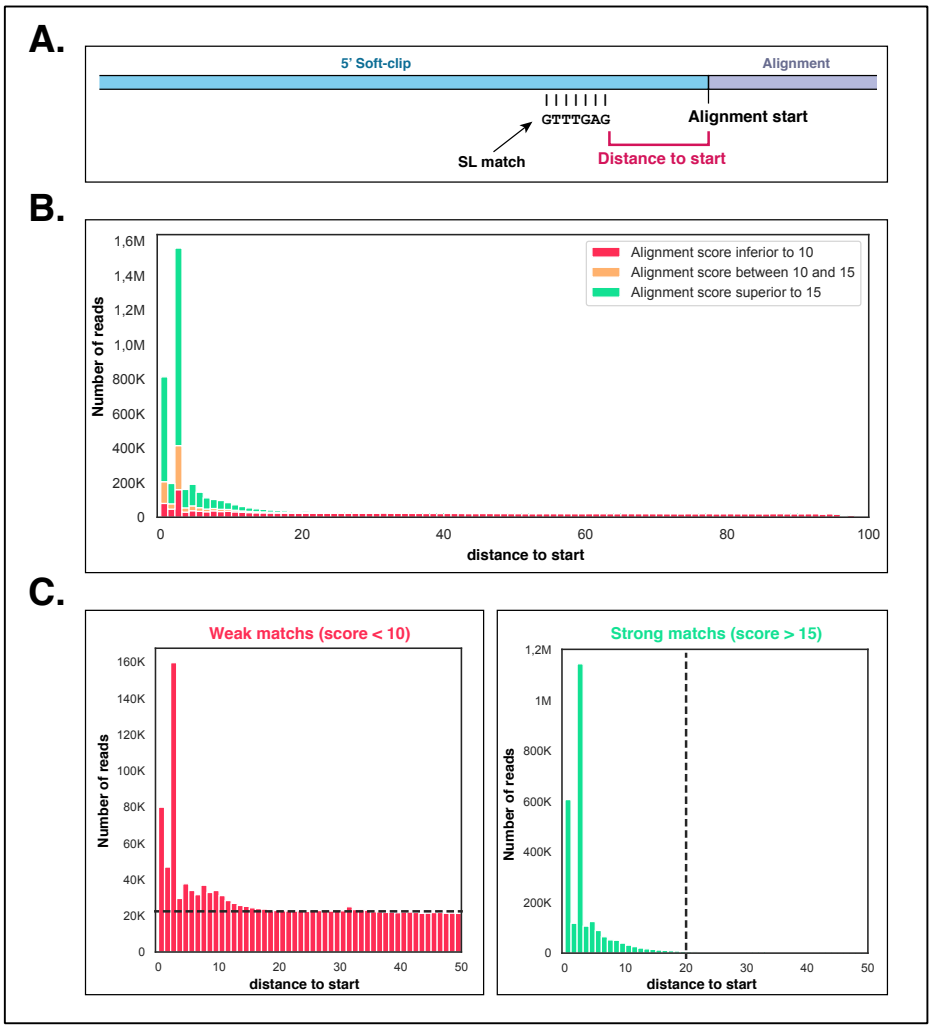


Figure 48 - Evaluation of distance to start for a given SL match. A) Schematic of how distance to start is measured for any SL match found in the soft-clip sequence. **B)** Distribution of distance to start in function of three categories of SL match: weak matches (score below 10), medium matches (score between 10 and 15) and strong matches (score above 15). **C)** Distance to start for weak (score below 10) and strong (score above 15) matches for a SL sequence.

From a total of ~8.3M reads mapping to protein coding genes, we managed to obtain a hit for a SL sequence in ~5.8M of them. Following filtering of low scoring hits, we finally accepted ~3.8M

hits, which represent 45% of all the reads. Out of the 17 886 genes detected in total, we could detect at least one read with a SL sequence in 13 306 different genes.

We then ordered our genes from the less expressed to the most expressed, based on the number of reads/gene, and then regrouped them in bins of one thousand genes. For each bin, we counted the mean number of read per gene and then we counted the number of genes for which at least one read had been detected with a SL sequence (**Figure 49**).

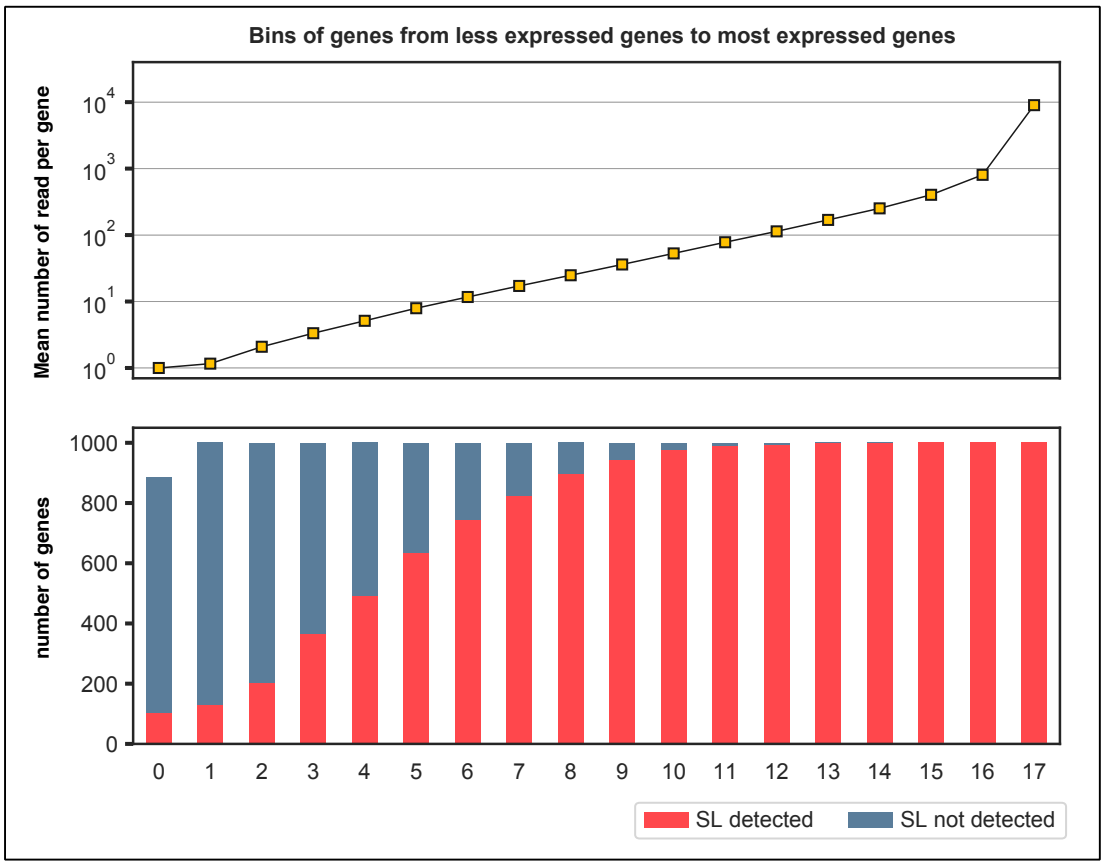


Figure 49 - Number of genes with an SL sequence detected from bins of 1000 genes ordered from the less expressed to the most expressed genes in our dataset. The top panel represents the mean number of gene detected in each bin, the bottom panel represent the number of genes detected with a SL sequence in each bin.

From this, we observed the 3000 most expressed genes (with a mean number of reads/gene per bin superior to 400 reads) had all been detected with a SL sequence. Furthermore, as the mean number of reads/gene per bin decrease, we detect fewer genes with a SL sequence. This result seems to indicate that our ability to detect a SL sequence attached to a gene's transcript is correlated to its level of expression in our datasets, hinting that most genes - if not all - are subjected to trans-splicing but that a lack of sensitivity in the detection of rarer transcripts might be hindering our ability to detect the phenomenon.

2.2.2 - Building an algorithm for classifying trans-spliced messengers

As confirmed by the meta-analysis of exon-junctions performed in the group a few years ago (Tourasse et al 2017), trans-splicing usually occurs at a predefined acceptor site just a few bases upstream of the start codon. However, trans-splicing is not an error-free mechanism and we can observe splice leader occurring at potentially spurious sites. Knowing this, we concluded that the identification of a read showing a splice leader sequence was not sufficient to infer the trans-splicing status of the gene it originates from. Consequently, we aimed at developing a method that would allow us to classify and characterize trans-splicing events for any gene detected in our sequencing experiments. The method was built around the identification of several reads with splice leader sequence but also on the presence of specific features that we deemed typical of trans-spliced genes.

a) Complementary evidences of trans-spliced messengers

Peak positions:

Since trans-splicing preferentially occurs at specific acceptor sites, we reasoned that full-length reads originating from the same trans-spliced gene would all generate alignments starting near the same genomic position. On the contrary, shorter reads or spurious trans-splicing events would tend to generate alignments starting randomly along the coding sequence of the gene.

Since some genes can produce different transcripts, we first used genomic alignments to extract their start positions as a set of genomic coordinates. However, when plotting the number of alignments per start positions (**Figure 50.B**), we realized that, even though we used a splice-aware aligner, mapping errors were frequent (**Figure 50.A**). This was mostly characterized by a significant number of alignments starting precisely at an exon junction. This behaviour could be characteristic of a new, unannotated, alternative promoter, but after inspection of the sequence upstream the start of the alignment, we could confirm part of the unmapped region of the read correspond to coding sequence of the gene, confirming the presence of mapping errors and not a new start.

Consequently, we decided to work with transcriptomic alignments. As the read are directly aligned onto known cDNA sequences, the aligner does not need to take introns into account, which makes mapping less complicated as the sequence needs to be aligned in one block. However, when performing transcriptomic alignments, the reference used becomes the transcript sequence, whereas in genome alignment the references used are the different chromosomes. Furthermore, in case of alternative promoters, different transcripts coming from the same gene can have completely different genomic positions. Therefore, I first converted transcript-based positions into genomic positions in order to work with the same reference for the different isoforms of a given gene (**Figure 50.A**). I then plotted the number of alignments per start positions and compared with the results seen with the genomic alignment (**Figure 50.B**).

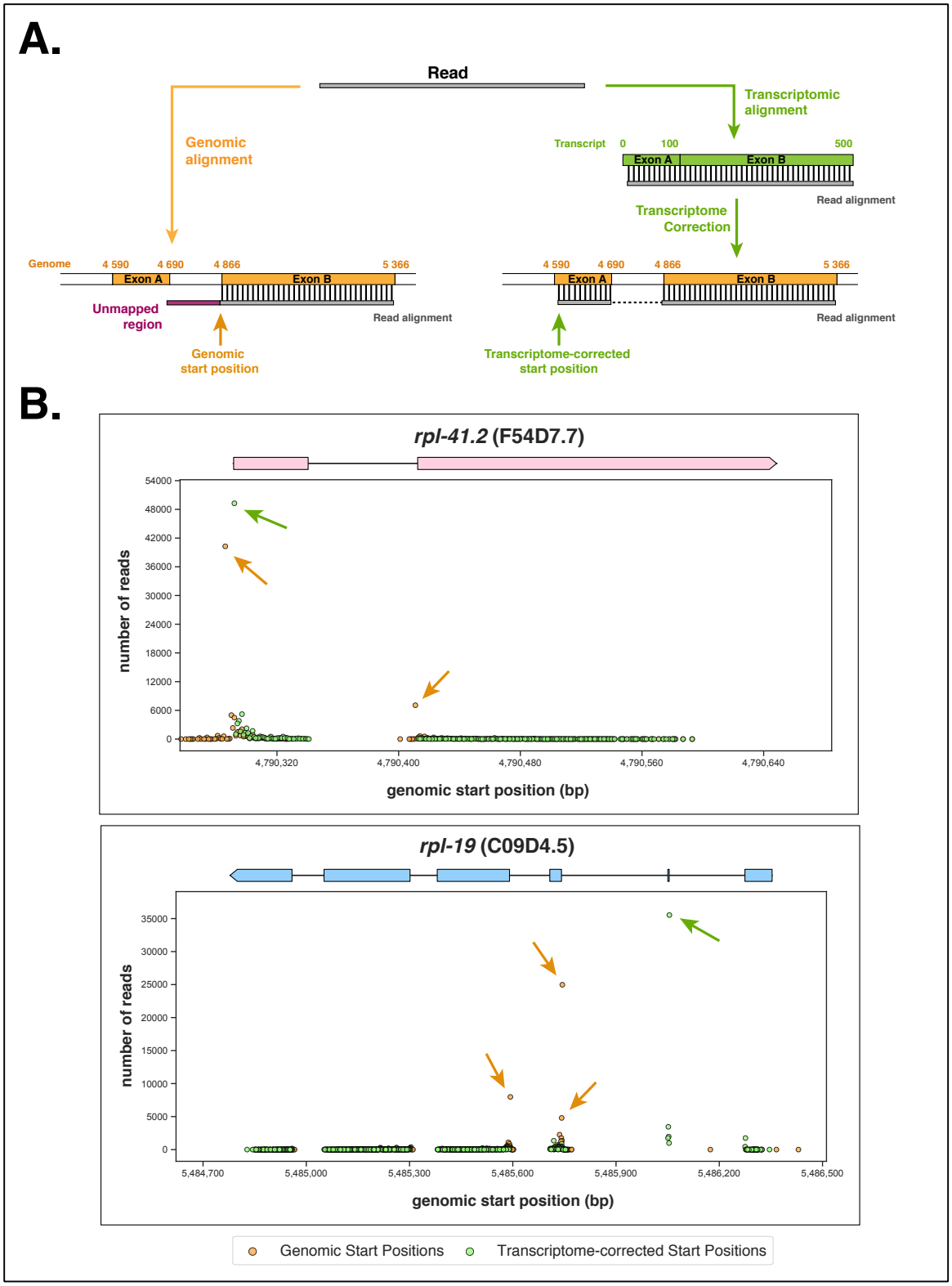


Figure 50 - Method for correcting genomic start coordinates. A) Schematic of how transcriptomic alignments are used for correcting genomic start positions: After transcriptome alignment (green), start and end positions (transcriptome-based) are converted to genomic coordinates (orange) in order to reduce mapping errors in splice-aware alignments. **B)** Example of corrected start positions for two different genes (*rpl-41.2* and *rpl-19*). Start positions extracted from

genomic alignments (orange) are compared with start positions extracted and converted from the transcriptome alignments (green).
By plotting the number of reads per start position, we could confirm that a significant number of reads generate alignments starting near the same genomic position, which imply we covered the entire 5' extremity of the messenger. The low-level background of alignments starting positions along the coding sequence are likely the result of "incomplete reads" (from broken mRNAs, incomplete sequencing, erroneous maturation events, etc.).

We then sought out to extract the main start positions (termed as "peak positions" in the rest of this manuscript) to differentiate which reads had generated those alignments and which had not. The determination of those positions was performed using a python script that I wrote myself. It works as follow:

For each gene, we pooled all of the reads coming from the different datasets and then evaluated their start positions (transcriptome-corrected) by using a pre-existing algorithm called "finding_peaks" from the well-established scipy library. Based on the documentation, this algorithm "finds all local maxima by simple comparison of neighbouring values".
Since genes are differentially expressed, we removed peak positions that are found lowly expressed by applying a filter based on the maximum amounts of reads found for the observed gene. We tested different threshold values, ranging from 0.1% to 10% of the total of reads found for any given gene (**Table 4**).

For each value tested, we evaluated:
- The number of genes for which we could detect at least one peak position.
- The number of genes for which we could not detect any.
- The mean number of peaks positions per gene.

From this, we observed that values of 2% and lower generated about the same results, the main differences being the number of peaks per genes that increase as the value tested decrease (2.86 peaks per gene for a threshold value of 0.1% versus 1.24 peaks per gene for a threshold value of 2%). However, using higher values (5 and 10%) significantly affected the number of genes detected, as expected from using more stringent criteria.

Based on those observations, we decided to use a default threshold value of 1%. This value was chosen (instead of using 2%) in order to allow to detect a slightly higher number of peaks.
Additionally, due to sequence noise and/or mapping errors, two very close positions can both produce peaks positions. Therefore, I decided to retain only consider the most expressed peak positions in a range of 20bp upstream and downstream.

Threshold	genes with peaks	genes without peaks	peaks per genes
0.1%	12950	4936	2.86
0.5%	12950	4936	2.25
1.0%	12949	4937	1.74
2.0%	12941	4945	1.24
5.0%	12730	5156	0.82
10.0%	11368	6518	0.65

Table 4 - Impact of using different threshold values for the detection of peak positions. Threshold values represent the minimum percentage of reads (from the observed gene) that needs to be present at a given start position in order to classify it as a "peak position". For each threshold value, we evaluated the number of genes for which at least a peak position was found, the number of genes for which no peak position was found and the mean number of peak positions per genes. In red is showed the final threshold value used in the rest of our analysis.

Distance to ATG codon:

Since finding peak positions is heavily dependent on the number of reads obtained for a given gene, we cannot always determine such positions for low expressed genes. In such case, we have decided to compare the distance between a start position and the nearest CDS start.
An alignment which starts near a known CDS start indicates the read it originates from is most likely full-length and, therefore, one might expect to observe a SL sequence on the 5' extremity if the messenger was indeed subjected to trans-splicing.

Presence of a long 5' soft-clip:

As shown in the first part of the results, long 5' soft-clip are caused by the presence of a SL sequence at the 5' extremity of the messenger during library preparation (see **section 2.1.2**). Hence, we used this artefact as an evidence supporting the trans-splicing of the messenger it originates from, even when it was not possible to detect the SL sequence itself.
To differentiate between "short' and "long" 5' soft-clips, we used the SL1 dataset. The interest of this dataset is the increased proportion of reads containing a 5' adapter instead of a hairpin, which makes it easier to determine their observed length. To do so, we measured the 5' soft-clip length of the reads in that library and plotted it as a histogram. As expected, we can see that most reads exhibit short soft-clip sequence of less than 100bp (~3.2M reads with short 5SC versus ~2.8M reads with long 5SC). Hence, we considered 5' soft-clip sequences longer than 100 bases as "long" 5' soft-clips and indicative of a trans-spliced messenger. On the contrary, reads with a 5SC sequence smaller than 100 bases were classified as "short".

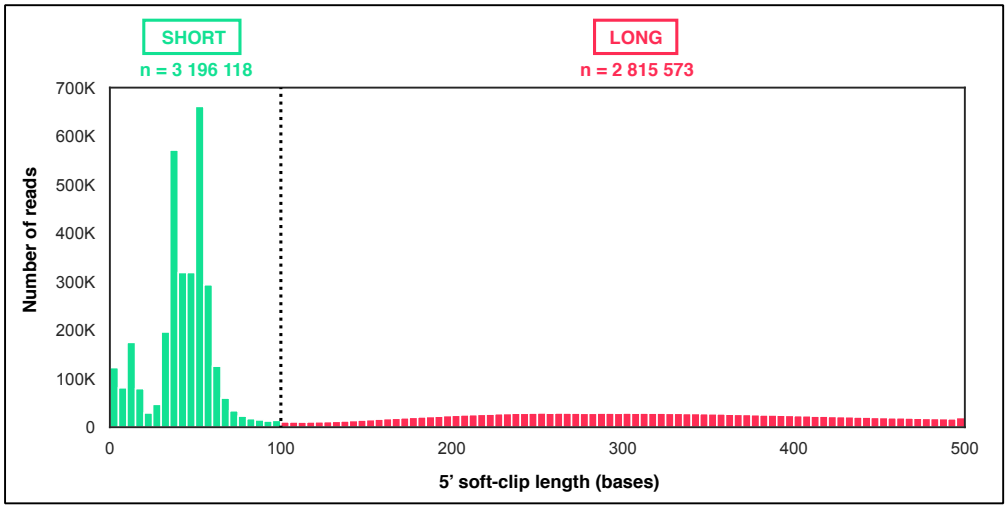


Figure 51 - Observed length of 5SC sequences when using the SL1 primer for 2nd strand synthesis. In this dataset, we observed an increased proportion of reads exhibiting the adapter sequence on their 5' extremity. We used those reads to determine the length of 5SC when adapters are present (in green) or when a SL hairpin is present (in red).

b) Building a decision tree for classifying trans-splicing events

By combining all the aforementioned features characteristic of trans-spliced messengers, we were able to build a decision tree to classify our level of confidence in the detection of a trans-splicing event for each of the gene (**Figure 52**).

As mentioned before, prior bulk analysis was performed for each gene in order to first detect the most represented start positions (peak positions). Then, for each gene, we analysed reads independently in order to detect if they start on peak positions or near a known ATG, if a SL sequence had been detected and if a long-soft-clip was present at the 5' extremity.

- Reads with shared starting positions (constituting a start "peak") and with either a SL sequence detected or a long 5SC were considered as strong evidence of trans-spliced messengers (TS+++).
- Reads starting near a known ATG position and with either a SL sequence detected or a long 5SC were considered as plausible trans-spliced messengers (TS++).
- Reads not starting on a peak position or near a ATG position, but for which we could detect a SL sequence or a long 5SC were considered as spurious trans-splicing events (TS+).
- Reads with a short 5SC but no SL detected were considered not trans-spliced (TS-).

After analysing each of the reads of a given gene, we were then able to pool those results to count the number of reads affected to each group (TS+++, TS++, etc.) in order to assess the general trans-splicing status of the observed gene.

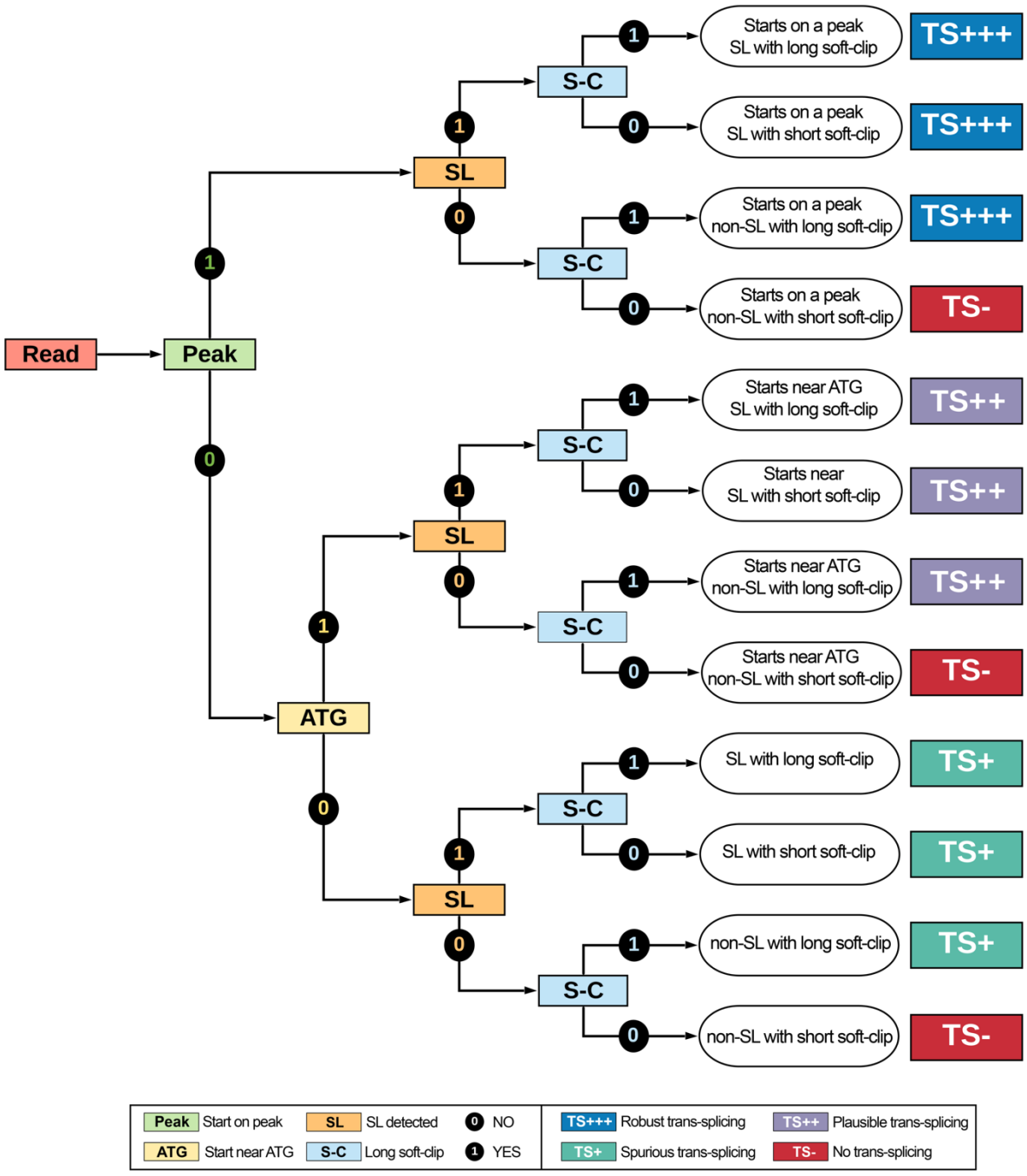


Figure 52 - Description of the decision tree used for classifying robustness of trans-splicing events. Prior bulk analysis of gene reads allows to determine peak positions. Each read is then evaluated independently. First, we look if the read start on a peak position. If not, we look instead if the read start near a known ATG position. Then we look if a SL sequence was detected and finally if a long 5' soft-clip is present. Reads are classified in four main groups based on the combination of the features found: "robust trans-splicing "(TS+++), "plausible trans-splicing" (TS++), "spurious trans-splicing" (TS+) and finally "no trans-splicing" (TS-).

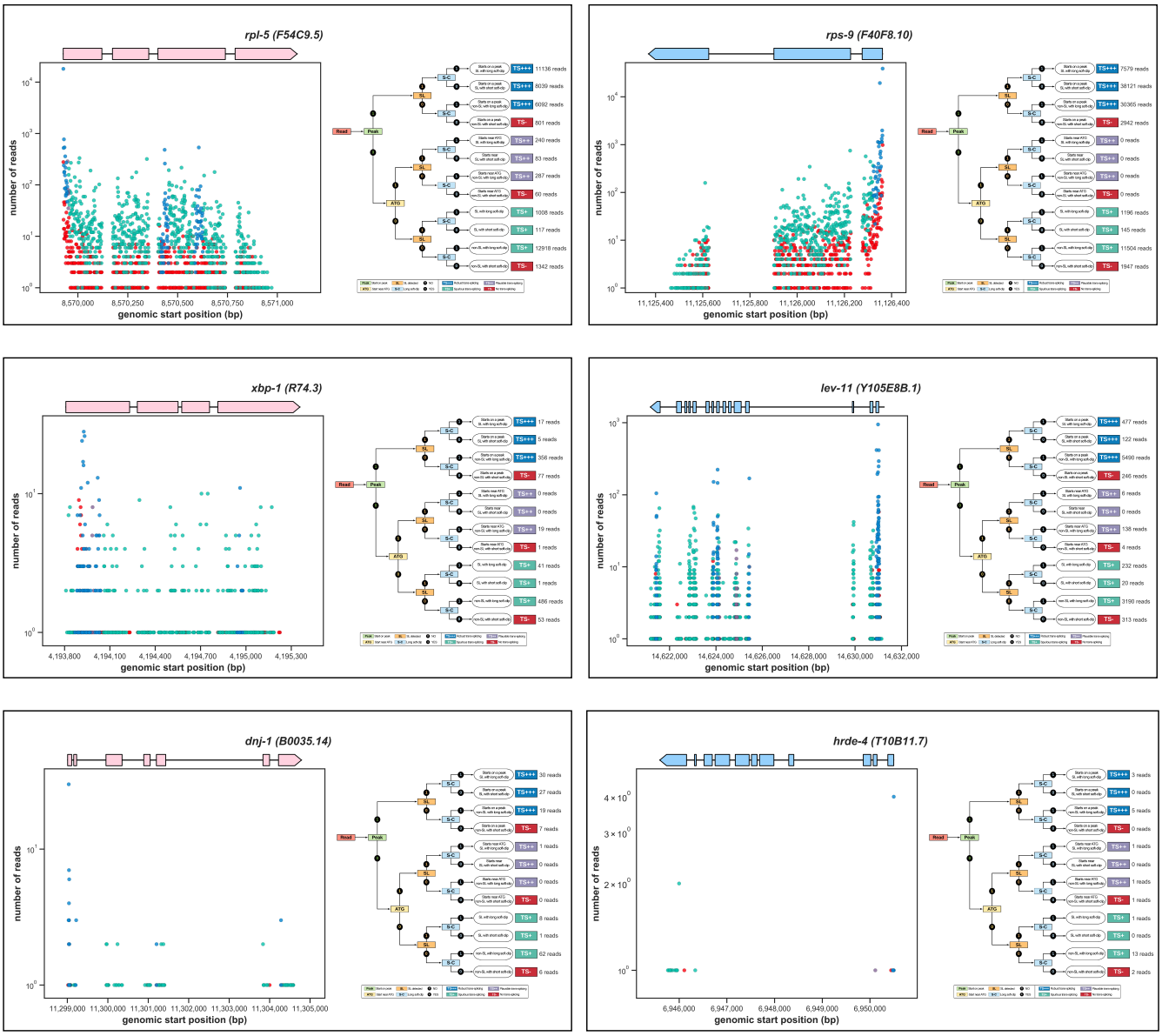


Figure 53 - Examples of trans-splicing analysis for six different genes. From top to bottom, left to right: *rpl-5*, *rps-9*, *xbp-1*, *lev-11*, *dnj-1* and *hrde-4*. For each panel, the number of reads found for each start position is showed on the left and results of the decision tree are shown on the right.

c) Filtering the obtained peak positions

Following classification of the reads, it was important to refine our results by finding positions for which trans-splicing had been found with robustness. To do so, in this section, we focused only on genes for which we could detect peak positions, since we consider those heavily represented positions as a major argument in favour of the identification of robust trans-splicing site.

Based on our decision tree, reads which start on a peak position can be split into four different classes depending on whether they exhibit a SL sequence and/or a long soft-clip. These classes correspond to the top four branches of the decision tree:

- Class I: SL detected and long soft-clip
- Class II: SL detected and short soft-clip
- Class III: no SL detected but long soft-clip
- Class IV: no SL detect and short soft-clip

First, we counted the percentage of reads attributed to each class for a given peak positions. We looked at different genes by plotting the results for each of their peak. Results for three different genes are showed in **Figure 54**.

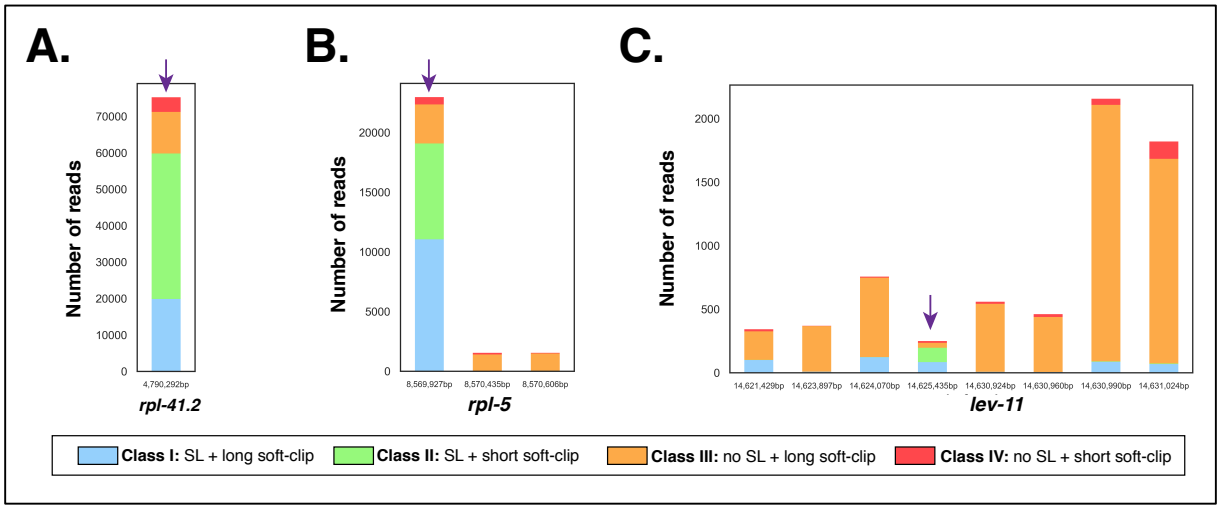


Figure 54 - Number of reads from each class for peak positions of different genes. Purple arrows indicate genuine trans-splicing sites

In **panel A**, we can observe gene *rpl-41.2* only has one peak position. Furthermore, we found a SL sequence for about 80% of the reads starting at this position (sum of class I and class II reads) which confirms this position as a genuine trans-splicing site.
For gene *rpl-5* we found three different peak positions (**panel B**), however only the first position shows a high level of SL sequence (~83%) while the two others only have about 3%, suggesting that only the first position correspond to a trans-splicing site. This observation is further strengthened by the fact that the first position is significantly more used than the two others (~23K

reads versus ~1.5K reads). When looking at the percentage of reads presenting a long soft-clip sequence, we observe ~60% of reads with a long soft-clip for the most represented position while the two other positions present more than 90% of reads with a long soft-clip, suggesting this value alone is not sufficient to discriminate trans-splicing sites.
Concerning gene *lev-11*, we initially detected eight different peak positions (**panel C**). Yet, just as in the previous example, only one of those seems to correspond to a trans-splicing site due to the high percentage of SL found (~79%). Unlike for *rpl-5*, this position is not the position for which we found the highest number of reads.

I tried to estimate if there is a characteristic repartition of read type that would be typical of a robust trans-splicing site. To do so, I selected all of the genes with a good coverage (more than 100 reads/gene) for which we had detected a single peak position. This gave us a list of 102 peak positions and, for each of those, I plotted the percentage of reads from each class (**Figure 55**). On this plot, we can most of the peaks positions exhibit between 10-30% of reads from class I, 40-60% of reads from class II, 10-20% from class III and less than 10% of class IV. If variability can still be observed between the different genes, this indicate that robust trans-splicing site present a characteristic repartition of read type. Nonetheless, out of this set of 102 robust sites, we detected two genes which showed a distribution of reads per class significantly different from the others, with a particularly high percentage of reads attributed to class IV (no SL detected and no long soft-clip).

In the first gene, *rpl-37*, we detect about 80% of reads from class IV at the peak position. Nonetheless we also observe ~8% of reads with a SL sequence, suggesting it is indeed a trans-splicing site. By inspecting the sequence around this position, we observed a genomic region that contains a sequence almost identical to the end of the SL1 sequence (16 bases in common, out of 22) (**Figure 56.A**). Since this sequence is present in the annotation of the genome, this resulted in the SL1 sequence being considered as part of the alignment and, therefore, it was not detected by our search algorithm since SL sequences are expected to be found in the 5' soft-clip.

For the second gene, *F46A8.7*, no SL sequence was detected and ~76% of the reads present a short soft-clip. However, the other 24% present a long soft-clip that is indicative of the presence of a SL. When we looked this gene in the wormbase database, we noticed its very small coding sequence (111bp). Because of the small size of its messenger, it is likely that 5' soft-clip sequence will rarely reach the size of 100bp that we estimated necessary for the classification of a "long" soft-clip. In this case our size threshold was too stringent to recognize the presence of a hairpin.
I also looked at the genomic region near the peak position in order to see if there could be a region similar to the SL sequence (**Figure 56.B**). We could only identify a very short motif of 5 bases resembling the end of the SL1 sequence that could affect our scoring system if the algorithm considers these as part of the aligned region rather than the 5SC.

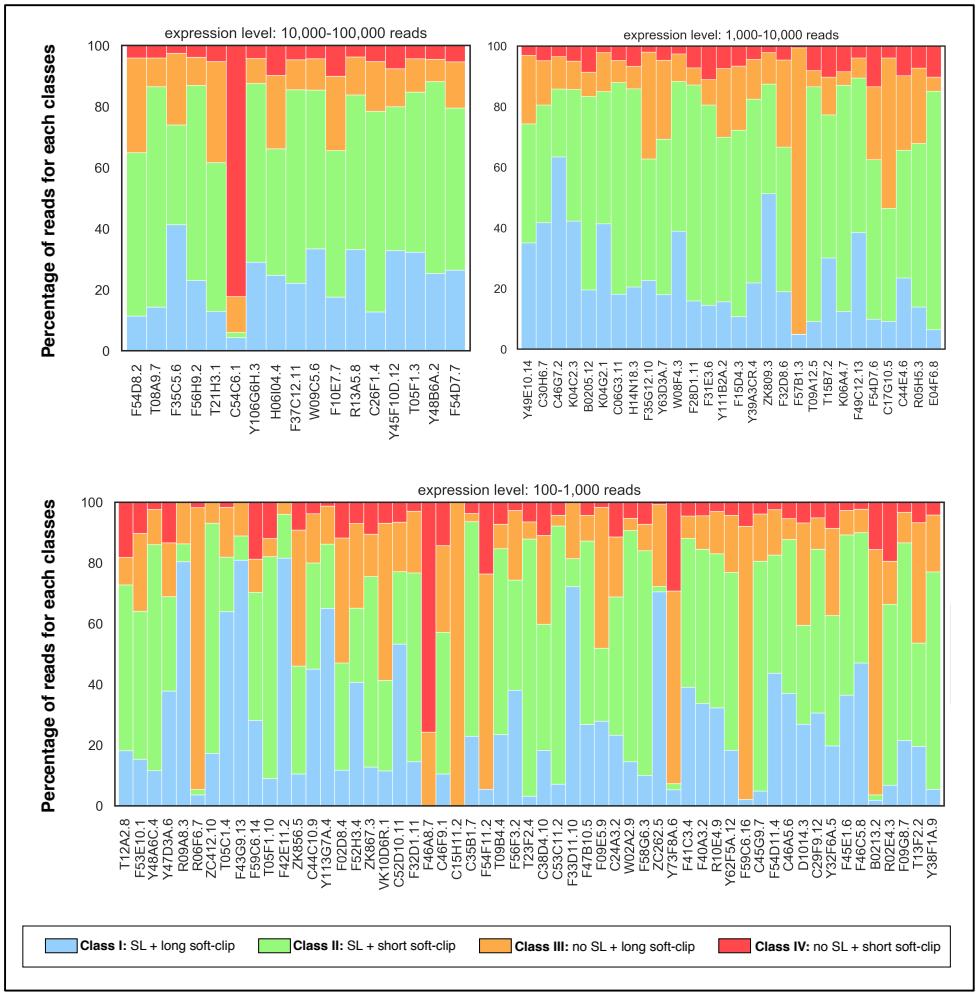


Figure 55 - Repartition of reads class for 102 genes considered robustly trans-spliced. Genes only present a single peak position and are split into three groups based on their general level of expression.

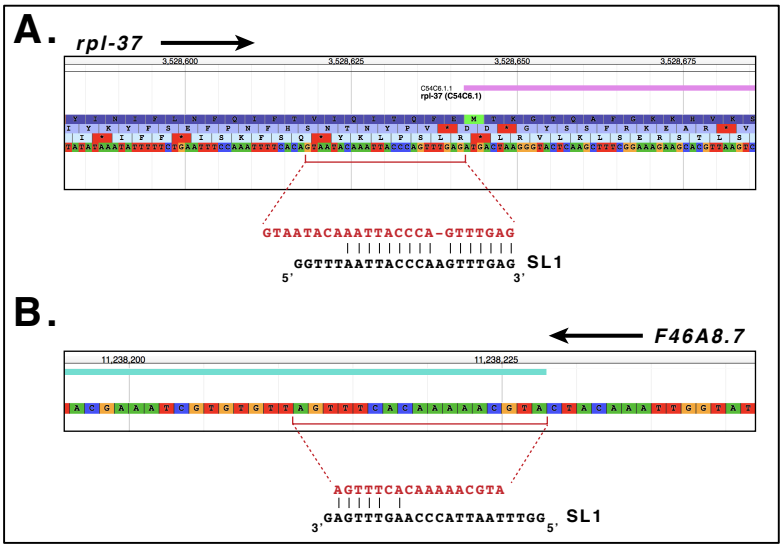


Figure 56 - Investigation of peaks positions for genes *rpl-37* and *F46A8.7*. The arrow next to the gene name represent its orientation on the genome. For each gene, we observed sequences resembling the end of the SL1 sequence just upstream the peak position.

With our list of 102 genes, we plotted the percentage of reads attributed to each class as a violin plot (**Figure 57**). We then used quartiles values to estimate a range of values that seemed representatives of what was generally observed at their peak positions. In order to have a range of values representative of the observed distribution, I chose to consider values situated between the first and third quartiles. However, when filtering peaks positions based on those values, we were only able to identify 603 different genes, which suggests the parameters we used might be too strict for accurately identifying most of the genes.

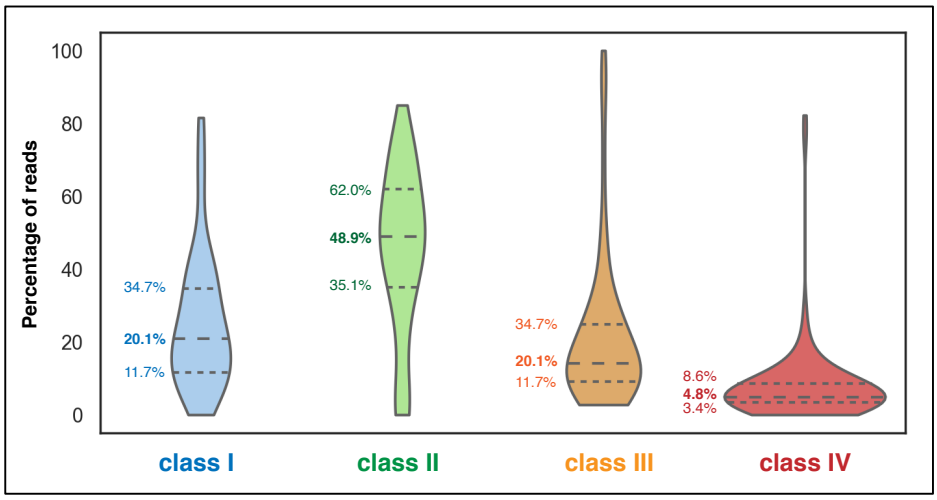


Figure 57 - Estimation of the repartition of reads class in a set of 102 genes robustly trans-spliced. For each class of read, 1st and 3rd quartiles values, along the mean value (bold), are indicated.

We then decided to look at a second criteria that might be indicative of a genuine trans-splicing site, which is the distance between a given peak position and the nearest annotated start codon. To do so, for each gene, I retrieved the genomic position of its known start codons and, for each peak position, I computed the distance with the nearest start. Since those positions are based on genomic coordinates, it is possible to observe large distances between an ATG and a peak position if there is an intron between them, however since trans-splicing is thought to happen in close proximity of start codon, we considered the influence of intronic sequences to be negligible for now.

We then took the list of 603 genes previously identified and looked and plotted the distribution of value for the nearest ATG codon (**Figure 58**). From this plot, we can observe that most of the peak positions are located just upstream an annotated start, with the large majority of them being situated less than 25bp upstream. This confirmed that our set represented indeed genuine SL sites and reinforced that we could use this distance in order to identify genuine trans-splicing sites.

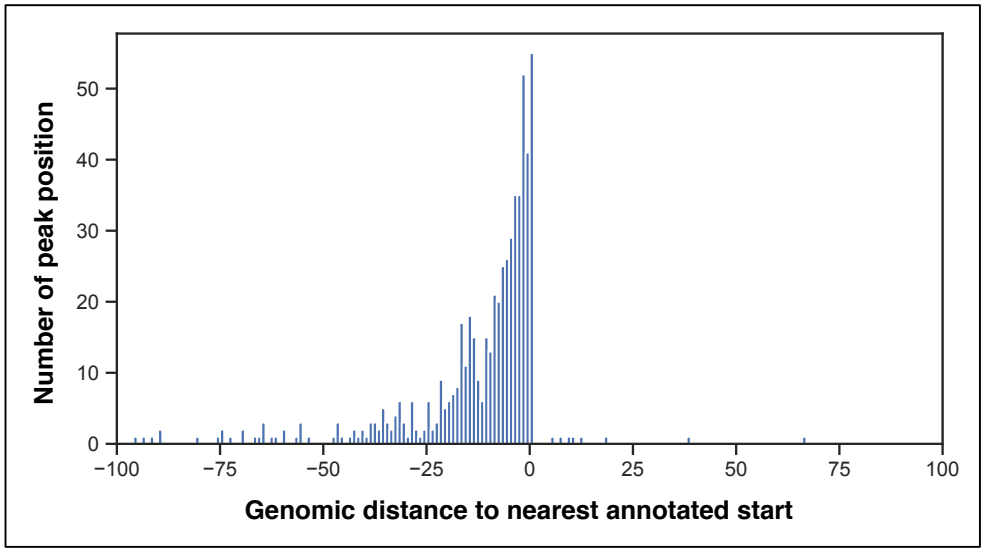


Figure 58 - Number of peaks positions found in close proximity of an annotated start codon. Start codon position is referred as 0. Negative values represent peak position located upstream their nearest start codon and positive values represents peak positions located downstream.

We then plotted the percentage of SL sequences detected at a given peak position in function of the distance to the nearest start (**Figure 59.A**). Once again, we observed that peaks positions with more SL sequences detected (above 40%) are generally found just upstream the start codon, while peak positions showing lesser percentages (below 20%) does not seem to be found at a specific location, suggesting they represent non-specific sites. However, such peaks positions presenting a low percentage of SL sequences detected still present a high percentage of reads with long soft-clips. This observation reinforces the idea that spurious SL sequences can be found at random places along the coding sequence of the messenger RNAs and, therefore, that trans-splicing is a more ubiquitous process than what was previously thought.

In order to see how this measure of distance to an annotated ATG start could help us in refining our results, we decided to measure the number of genes and the number of peaks positions detected in function of this distance (**Figure 59.B**). We evaluated from -100bp to +150bp (0 being the position of the annotated start codon), and we found that a majority of genes had a peak position located upstream of a start codon (up to +1bp).
When considering peak positions situated downstream of their nearest start codon, the number of new genes identified did not increase a lot while the number of peaks positions increased significantly, suggesting those peak positions are not genuine trans-splicing sites but rather are the result of using a low threshold for the detection of peak positions.

With this observation, we decided to retain genes for which a peak position situated at +1bp or less from their annotated start. However, since the presence of SL sequences at a given peak is a

strong indicator in favour or robust trans-splicing sites, we decided to also retain any peak position for which more than 20% of reads were detected with a SL sequence (class I and II).

Out of the 31 710 peak positions initially identified, we finally retained 15 654 of those with confidence, for a total of 12 082 unique genes. This left us with only 867 genes for which we could not detect a robust trans-splicing with high confidence.

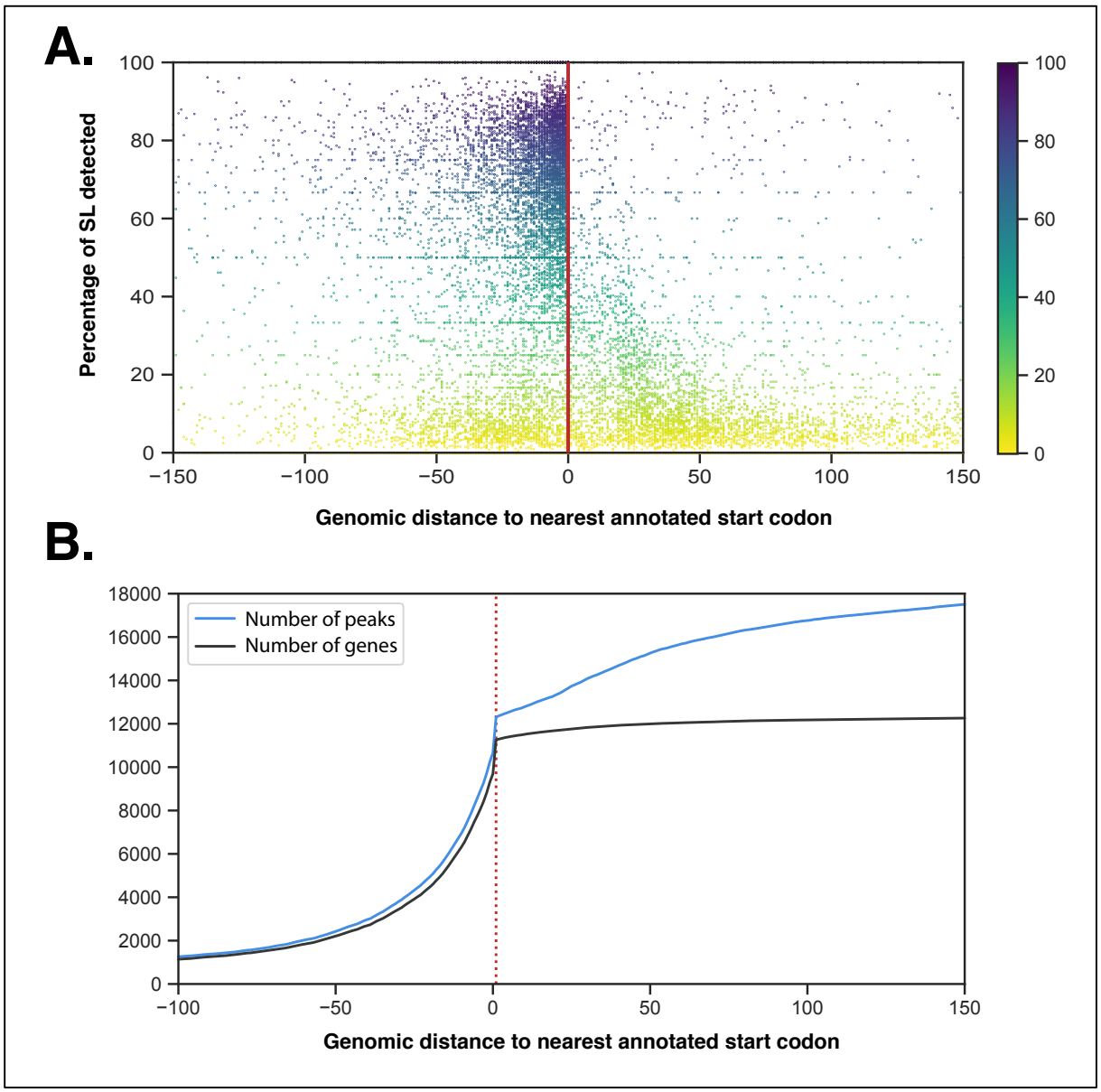


Figure 59 - Filtering of peak positions by their proximity to an annotated codon start. A) Percentage of SL reads detected (class I and class II reads) in function of their distance to the nearest codon start. **B)** Number of genes and peak positions retained when filtering in function of their distance to the nearest start codon. Dotted line in red represent the chosen value (+1bp).

2.2.2 - Is trans-splicing a pervasive mechanism?

Following the determination of robust trans-splicing sites, we looked at the level of expression of the different categories of genes as determined initially by the decision tree. Robust trans-splicing genes were split in two groups following peak positions filtering. Genes for which we retained at least one peak positions were considered as "confirmed" robust trans-splicing site and the others as "non-confirmed".

For each category of genes, we plotted their number of reads as a boxplot (**Figure 60**). This plot allows to observe that genes with higher level of expression are more likely to be detected as robustly trans-spliced, whereas genes for which we have weak evidences are lowly expressed (less than 10 reads/gene). Furthermore, genes for which we could not provide evidences of trans-splicing activity are only detected by 1 or 2 reads, suggesting once again that our ability to detect trans-splicing events with confidence is primarily limited by gene coverage: the more data we gather on a given gene, the more likely we are to detect trans-splicing.

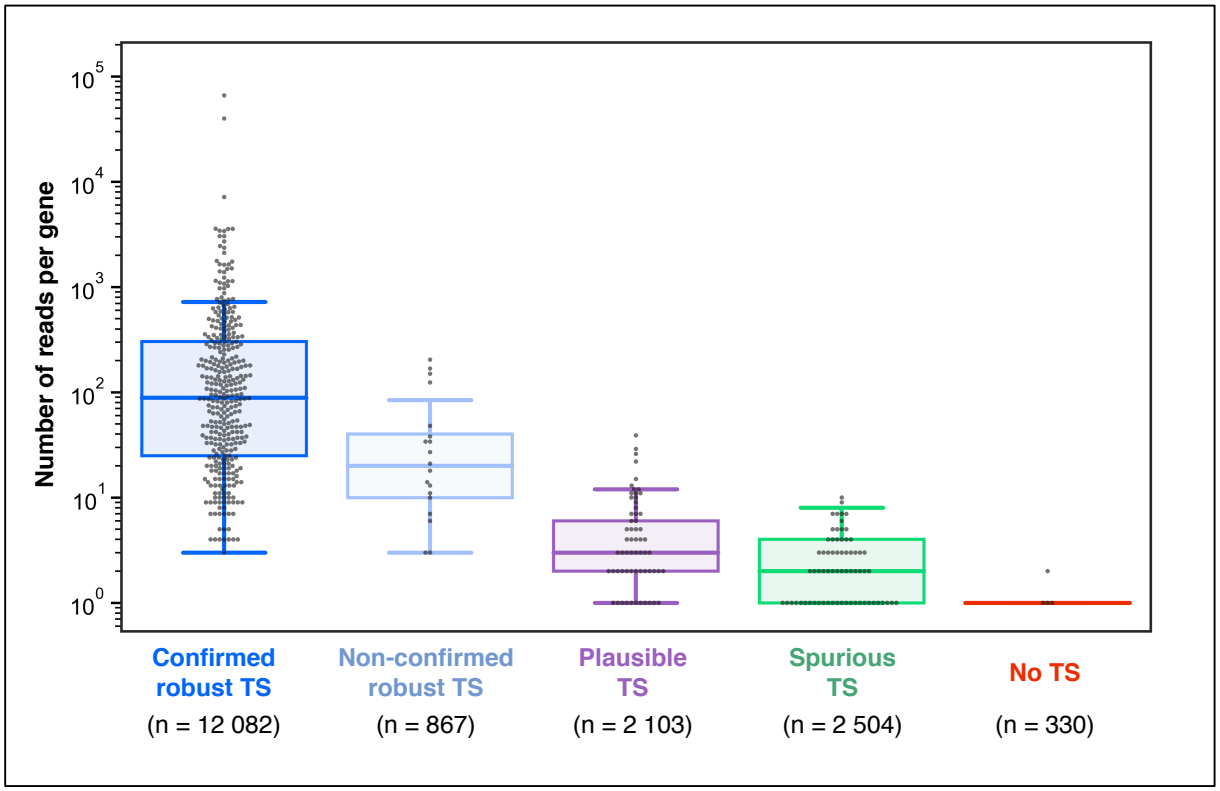


Figure 60 - Gene's expression level within each category. Boxplot represent the observed level of expression for each class of gene. On top of each boxplot, we added gene expression level (black dots) for a subgroup of 1500 genes randomly chosen in our dataset. The total number of genes in each group is annotated below each category name.

We then decided to cross those results with the meta-analysis performed by Tourasse *et al*, 2017 (**Figure 61**). For each category of genes, we measured how many of them had also been found in their analysis. Genes detected with high confidence in our dataset were almost all found in the meta-analysis (93%), however, this number decrease for population of genes that are generally less expressed. Out of the 330 genes for which we have no evidence, 201 genes (60%) have been found trans-spliced by the meta-analysis. This suggests once again that the level of expression of a given gene is the main limiting factor in the elucidation of its trans-splicing status.

When we cross the results of the two analyses (ONT sequencing + meta-analysis), we can see that a large majority of genes are detected in both dataset (15 946 genes). 2 263 genes are uniquely found in the meta-analysis, but our own Nanopore sequencing also provides evidence for 2 060 genes that were previously not detected. As new studies are performed, and as the resolution increase, we detect more genes being subjected to trans-splicing, which suggests that the mechanism is more prevalent than previously thought and indeed ubiquitously used for all RNA polymerase II transcripts.

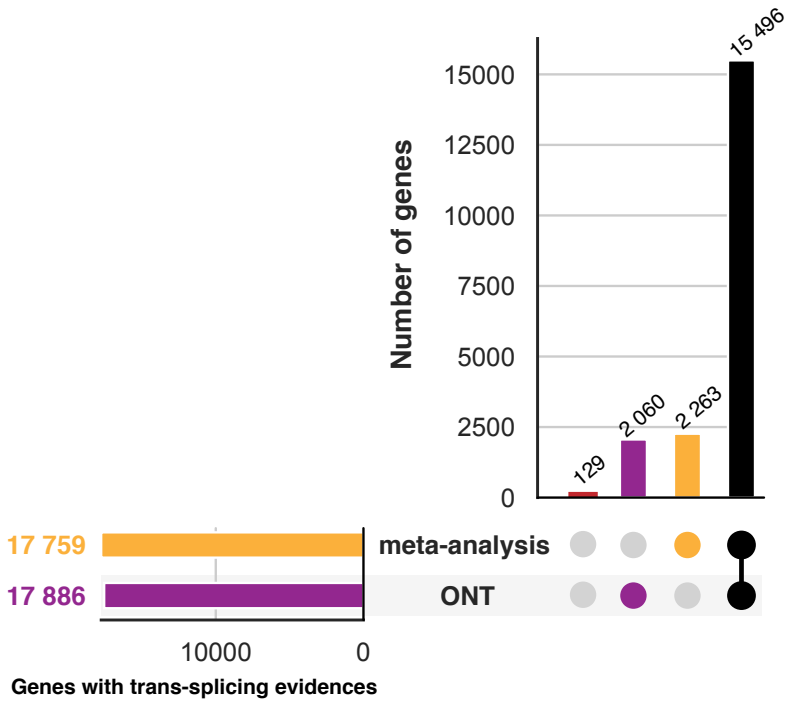


Figure 61 - Gene sets comparison between our dataset (ONT) and the meta-analysis of exon junctions generated by Tourasse *et al* (meta-analysis). An upset plot was generated to visualize gene sets intersections. In black are the genes for which trans-splicing was detected in both datasets, in purple genes for which it was detected only in the ONT dataset and in orange the genes for which it was detected only in the meta-analysis. Genes in red were not found to be trans-spliced in any of the datasets.

2.2.3 - *lev-11* gene shows differentially trans-spliced populations of mRNAs

During this project, we uncovered a particular behaviour regarding the trans-splicing of gene *lev-11*. As quickly mentioned in **section 2.2.1** (see **Figure 54**), we confirmed for this gene a robust trans-splicing site that did not correspond to the most represented peak positions.

After investigation of this gene, we noticed the presence of two alternative promoters: a distal promoter who is responsible for the production of long isoforms and a proximal promoter responsible for the generation of short isoforms. Genomic coordinates of the peak positions for which we detected robust trans-splicing correspond to the position of proximal promoter (**Figure 62**).

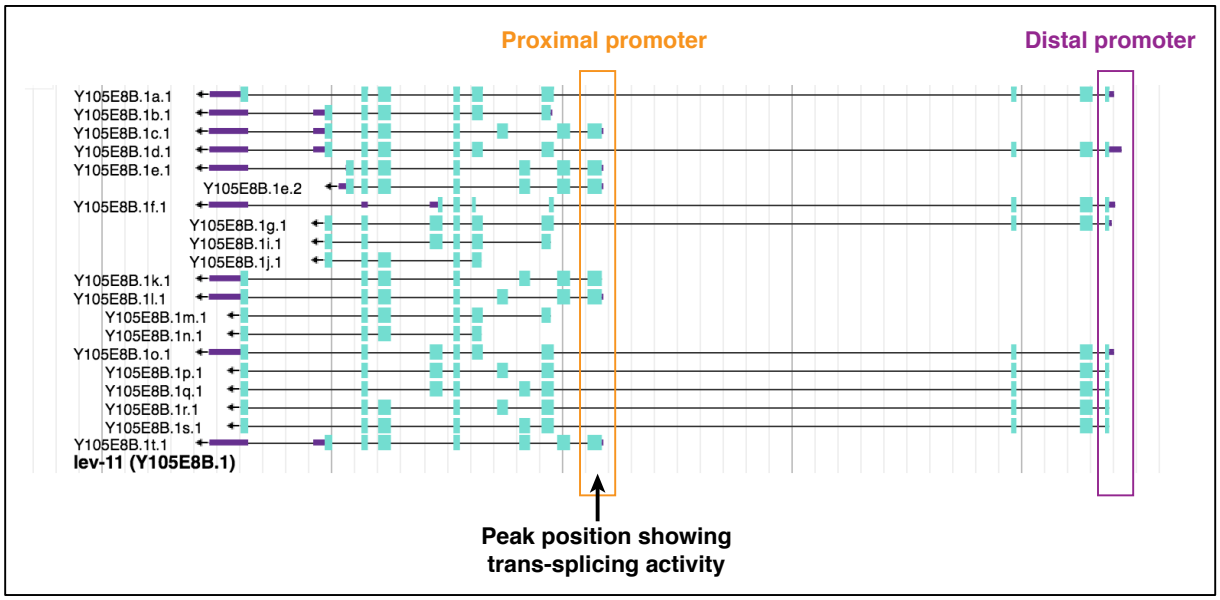


Figure 62 - *lev-11* present alternative promoter sequences. Proximal promoter position is shown in orange and distal promoter position in purple. The peak position for which we detected trans-splicing activity is indicated by the black arrow.

After comparison with the result of the meta-analysis of exon junction (Tourasse et al, 2017), we observed that long isoforms are not found associated with any specific SL sequence, however the short isoforms are found associated with a SL1 sequence. Furthermore, based on their quantitative analysis of exon usage, long isoforms are preferentially expressed in *C. elegans* (80-90% of *lev-11* transcripts) (**Figure 63**). Therefore, this difference of expression between the isoforms produced by the two promoters explains the difference of start position usage that we observed in our previous analysis.

To confirm this behaviour, we compared the results obtained from SSP [polyA] direct-cDNA experiments with the SL1 library generated by RT-PCR when we performed the comparative

analysis between the different sequencing kits. The rationale behind the use of this RT-PCR dataset is that, since reads were amplified using a SL1-specific RT-PCR, we expect to find only the short isoforms if the proximal promoter is not associated with SL1 trans-splicing.

After comparing and measuring the number of reads attributed to either forms (long or short isoforms), we were able to report that only the short isoform is found in the SL1 RT-PCR dataset (616 short isoforms out of 616 reads), while the SSP [polyA] dataset provided ratios that are consistent with the observed exon usage found in the meta-analysis: 254 (9%) and 2537 (91%) reads for the short and long isoforms respectively.

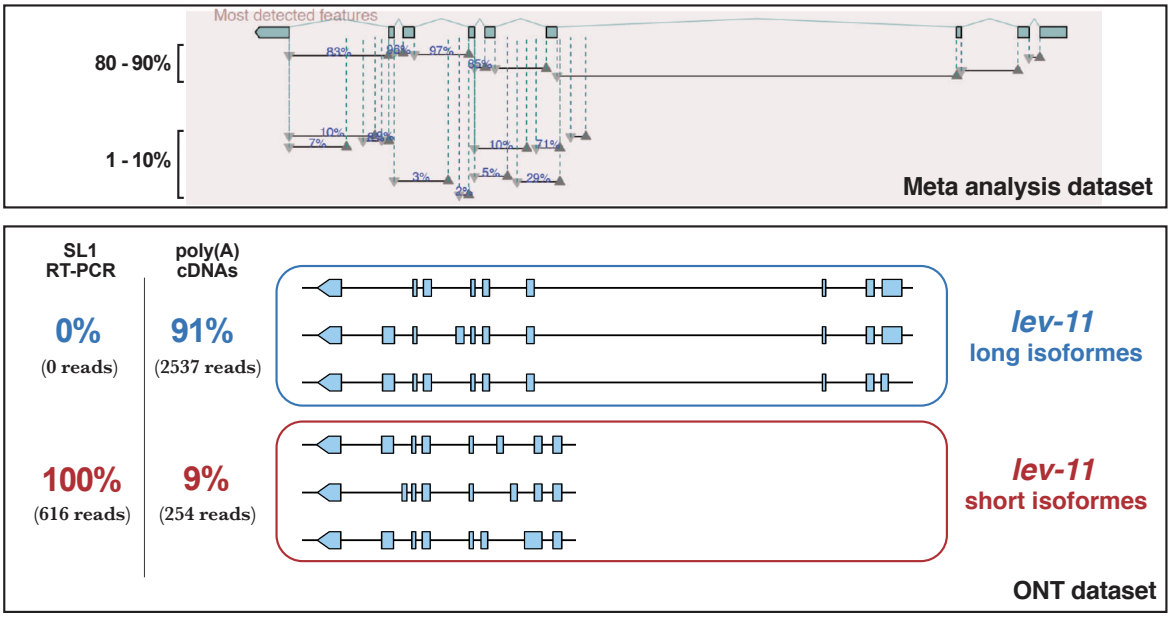


Figure 63 - Comparison between the results of the meta-analysis of exon usage in *C. elegans* with the ratio of the different isoforms of *lev-11* detected in different sequencing experiments

This observation demonstrates that trans-splicing specificity can be promoter-dependent.
However, if we could not identify any SL1 or SL2 sequence on the 5' extremity of *lev-11* long isoforms, we still observed long 5' soft-clip sequences, suggesting the presence of a SL hairpin.

Thus, we then decided to extract the end of the 5SC sequence for the peak positions that is located near the proximal promoter. When looking at the sequence just upstream the start of the alignment, we observed the presence of a 18nt sequence that did not correspond to any genomic region of the of *lev-11*. We then performed a BLAST analysis for this sequence and discovered a match for two non-coding RNAs (*Y59A8B.37* and *Y51H7C.37*) for which no function has been described so far. This lead us to postulate those ncRNAs might serves as a new Splice Leader sequence, however after looking at the ability of this sequence to form a hairpin structure at the end of *lev-11* long isoforms - and therefore to produce long 5SC as seen in those reads - we observed that this

18nt sequence is entirely complementary of the 5' extremity of the gene, suggesting that the end of *lev-11* extremity might able to form hairpins on its own.

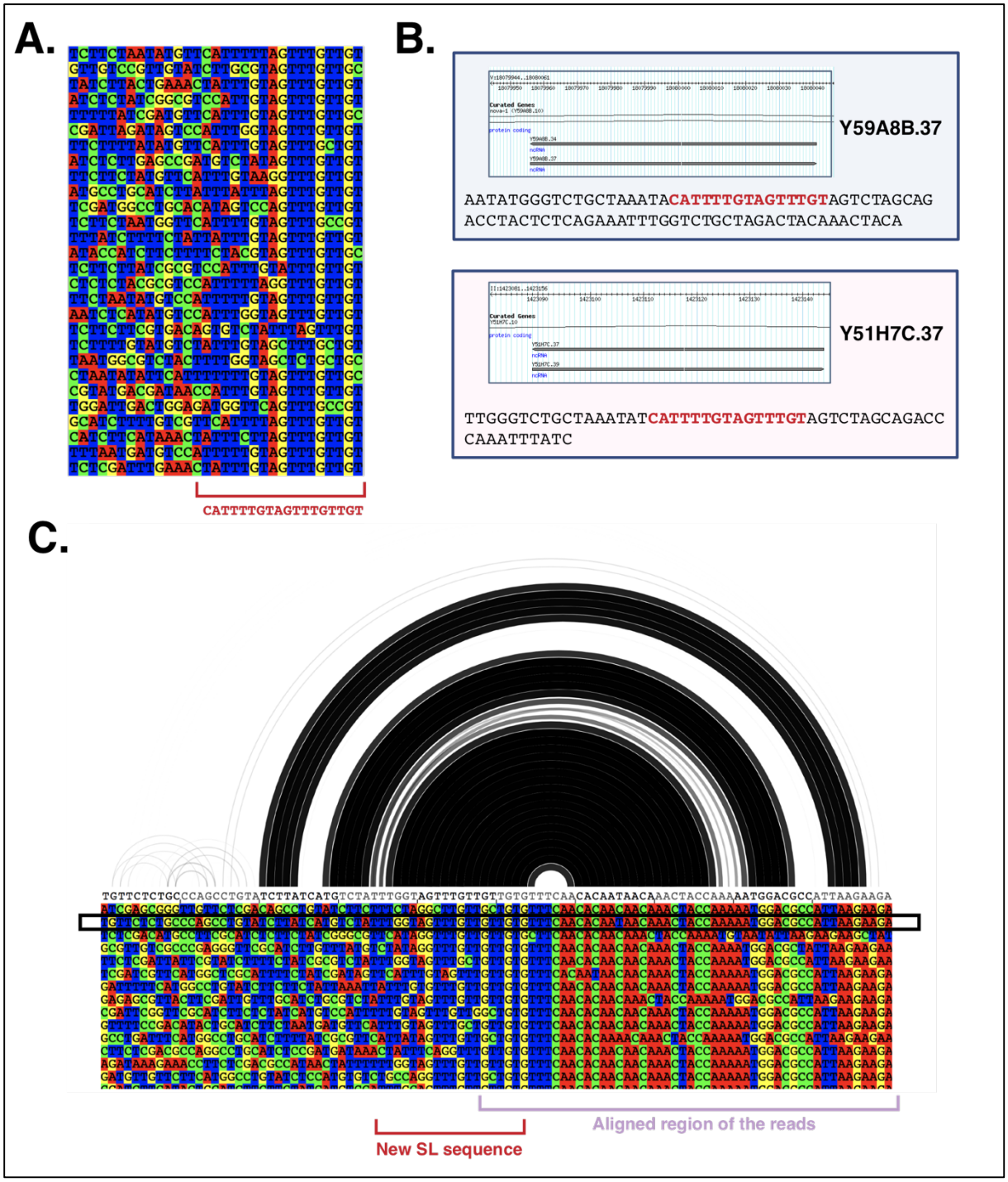


Figure 64 - ***lev-11* long isoforms present a hairpin structure on their 5' extremity**. **A)** Identification of a 18nt sequence at the 5' extremity of the read. **B)** The 18nt sequence match with two non-coding RNAs present on *C. elegans* genome. **C)** The 18nt structure is complementary of *lev-11* long isoforms 5' extremity.

2.3 - Tissue-specific transcriptome analysis in *C. elegans*

Tissue dissections in mice or humans has allowed to generate large-scale tissue maps for better understanding the differences between different cell-types at the level of both RNAs and Proteins. However, performing dissections in *Caenorhabditis elegans* is a challenging task due to the small size of the animal. Despite being the first metazoan whose genome was fully sequenced, most genes of *C. elegans* have not been analysed at a tissue-specific resolution and our understanding of the link between cell type and gene expression remains incomplete.

However, in the past decades, many efforts have been made toward the development of a comprehensive tissue-specific transcriptome map. Gene expression in early embryos has been extensively studied and a method for dissecting gonads has been developed to allow the study of the germline (Seydoux and Fire 1994). Furthermore, researchers have also been using Fluorescence-Activated Cell Sorting (FACS) techniques, by expressing fluorescent markers in specific tissue, which allowed to make considerable progress in that area. Today, new methods are still being developed, as we can see with the publication of the Slice-seq method (Ebbing et al. 2018), which involves to perform single-cell sequencing on thin animal slices, which contributed to better understanding spatial transcriptomics in worms. However, those different techniques require complex protocols and can be very expensive to set up.

In this section is described our efforts at combining a recent technique which allow us to easily tag RNAs in a tissue of interest along with the advantages of third generation sequencing.

2.3.1 - Construction of vectors for tissue-specific expression

a) Experimental design

As seen in previous sections, *C. elegans* mRNAs undergo a trans-splicing process as they are matured, resulting in the addition of a SL sequence replacing their 5' untranslated region. This sequence being produced from a SL gene, it was shown that it is possible to create a transgene carrying a tagged SL sequence that will be efficiently trans-spliced onto mRNAs (Ma et al, 2016). Hence, by placing this transgene under the control of a tissue-specific promoter, this allows for specifically tagging the messenger RNAs expressed in specific sets of cells (**Figure 65.A**).

The authors used the presence of the tag for performing tissue-specific isolations from total RNAs extractions, allowing us to retrieve only the RNAs of interest (**Figure 65.B**). Combined with the use of new sequencing technologies, we reasoned that not only should we be able to assess the splicing status of messenger RNAs in a given tissue, but also to quantify their level, enabling us to characterize tissue-specific expression with a never before reached resolution.

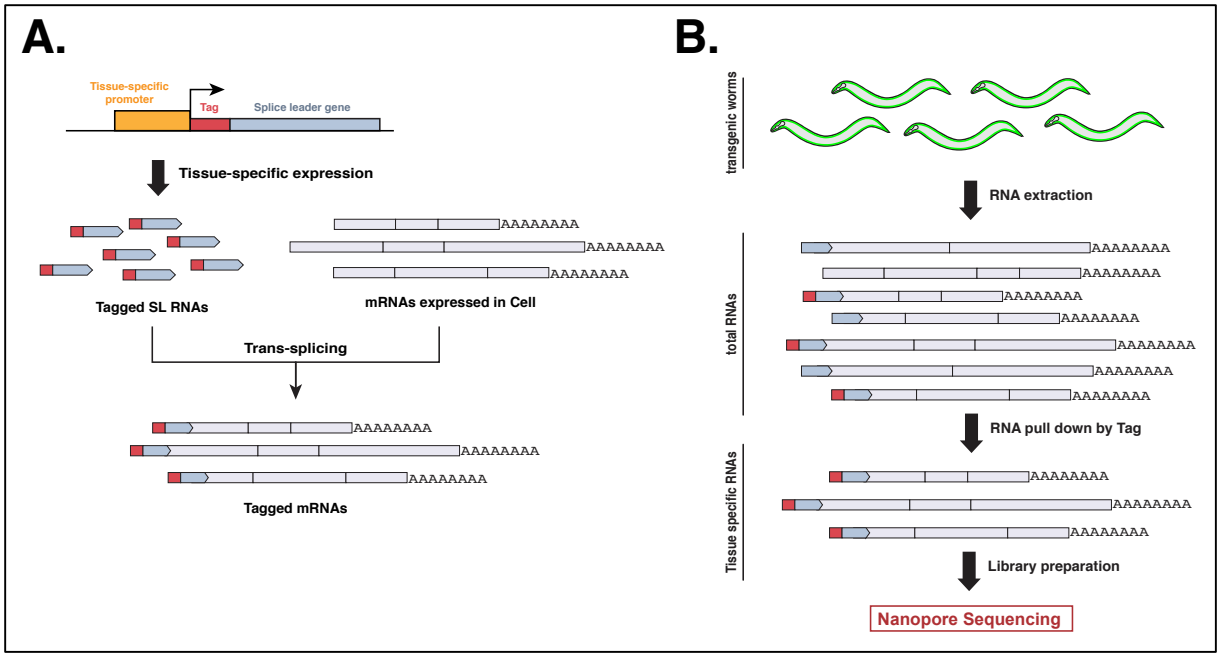


Figure 65 - Isolation and sequencing of tissue-specific mRNAs. A) Schematic of the method (Ma et al, 2016). A tagged SL sequence is put under the control of a tissue-specific promoter. mRNAs present in the same tissue as those modified SL will have a chance to be trans-spliced with it. **B)** Following total RNAs extractions from transgenic worms carrying the Tag::SL1 construct as an extrachromosomal array, the RNAs of interest are pull-down via their Tag sequence and subsequently used for nanopore sequencing.

To this end, we have selected a list of promoters of interest to drive specific expression of the tagged Splice-Leader transgene in various tissues of interest. First, we used the promoterome database (available at: http://worfdb.dfci.harvard.edu/promoteromedb/) for selecting the initial sequence of each promoter (Dupuy et al. 2004). Then we adjusted the length of the sequence based on newest publications - when available - in order to work with sequences of different sizes, allowing for rapidly screening for their insertion in new transgenes by using PCR amplifications. Furthermore, we decided to add a 4 letters barcode, between the tag and splice leader sequence, specific for each promoter. This was done for recognizing tagged RNAs coming from a specific tissue, allowing to expressed different constructs within the same organism or to multiplex different RNAs libraries in one single sequencing experiment.

As a trial set, I have selected three types of tissues of interest: muscles, neurons and hypodermis. The rationale behind this choice was to be able to analyse tissues that are very different from each other and that carry different functions. Furthermore, muscles and neurons are interesting tissues to analyse because subtypes of tissues (like body wall muscles and pharyngeal muscles) present different expression. This strategy could allow us to comprehensively catalogue differences at the level of expression between tissues that are either closely related or more different to each other. Other promoters, expressed in only a few cells, are also interesting because they will allow us to

test the sensitivity of our approach to determine how close we can get to single cell transcriptome characterization.

Promoter	Promoterome size	Final size	Promoter expression	Cellular type	Barcode
myo-3	2000	1607	Body wall muscle		-
myo-2	979	976	Pharyngeal muscle	**Muscles**	**CGTC**
hlh-8	2000	1187	Vulval muscle		**CTGC**
unc-119	2000	2200	pan-neuronal expression		**AGAT**
rgef-1	2600	2600	pan-neuronal expression		**AGGA**
ida-1	303	303	ALA, VC, HSN, and PHC neurons		**AGAG**
sng-1	2000	2000	dorsal and ventral nerve chord + anterior nerve ring		**AGTA**
mec-7	718	718	touch-receptor neurons (6 cells)	**Neurons**	**TCGG**
mec-8	2000	1630	mechanosensory neurons		**TCAA**
mec-2	2000	2000	mechanosensory neurons		**TCTT**
gcy-7	1446	1446	ASEL only		**ATAC**
gcy-5	2000	3000	ASER only		**ATAG**
unc-52	2000	1809	extracellular matrix between mucle and hypodermis	**Hypodermis**	**GACT**

Table 5 - List of promoter selected for driving expression in specific tissues. Specific 4 bases barcodes are introduced between the Tag and the SL sequence to discriminate tagged RNAs expressed in different tissues.

b) Plasmid constructions

For each promoter, two sets of vectors were generated:

- One vector expressing the Tag::SL1 construct for tissue-specific tagging
- One vector expressing GFP for validating the transcriptional activity of our promoter and quickly recognizing the different strains under fluorescent light.

The first vector was derived from the vector used in the study of *Ma et al.* This construct, expressing a Tag::SL1 sequence under the control of the *myo-3* promoter (body wall muscle expression) was send to us by the team of Dr. Xiao Liu (Tsinghua University, Beijing).
The second construct was derived from a plasmid construct already used in the lab expressing GFP under the control of a *myo-2* promoter.

For replacing the promoter of the initial vector with our own promoters, we used an insertion method called AQUA-cloning (Beyer et al. 2015). This method permits the insertion of an insert inside a vector by homologous recombination by using the innate *in vivo* activity of *E. coli*. To this end, we amplified the fragment of interest by using primers with a 32bp tail that is complementary from the sequence of the linearized vectors. This overlap between the two fragments allows for insertion of the fragment by homologous recombination after transformation inside competent cells (**Figure 66**).

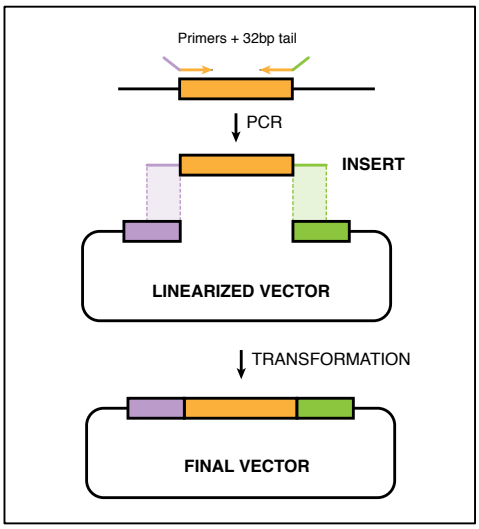


Figure 66 - General AQUA-cloning method. The insert is generated by PCR using primers with tails complementary from the linearized vector. After transformation in competent cells, the fragment is inserted into the vector by homologous recombination using the in vivo activity of the cell.

For the insertion of the promoter in the SL1 vector, we linearized the vector by PCR using a reverse primer hybridizing upstream the start of the promoter (backbone region of the plasmid) and a forward primer hybridizing at the start of the SL1 sequence. On the forward primer, we added a 32bp tail containing the Tag sequence and the 4 letters barcode corresponding to the promoter we want to insert. The promoter itself was amplified by PCR from genomic DNA using primers with 32bp tails complementary of the linearized vector (**67 - left panel**).

Concerning the creation of transcriptional reporters, we designed primers to linearize a Pmyo2::GFP plasmid in a way that permitted to remove the Pmyo-2 sequence and we added tails on the extremity of each primer to allow insertion of the promoter sequence already amplified. This allowed us to reduce the amount of PCR amplifications to perform. In this construct, the tail sequences were considered as "recombination sites", flanking the promoter sequence. This sequence containing a unique restriction not previously found in the backbone of the plasmid, it also allowed us to verify the insertion by enzymatic digestion (**Figure 67 - right panel**)

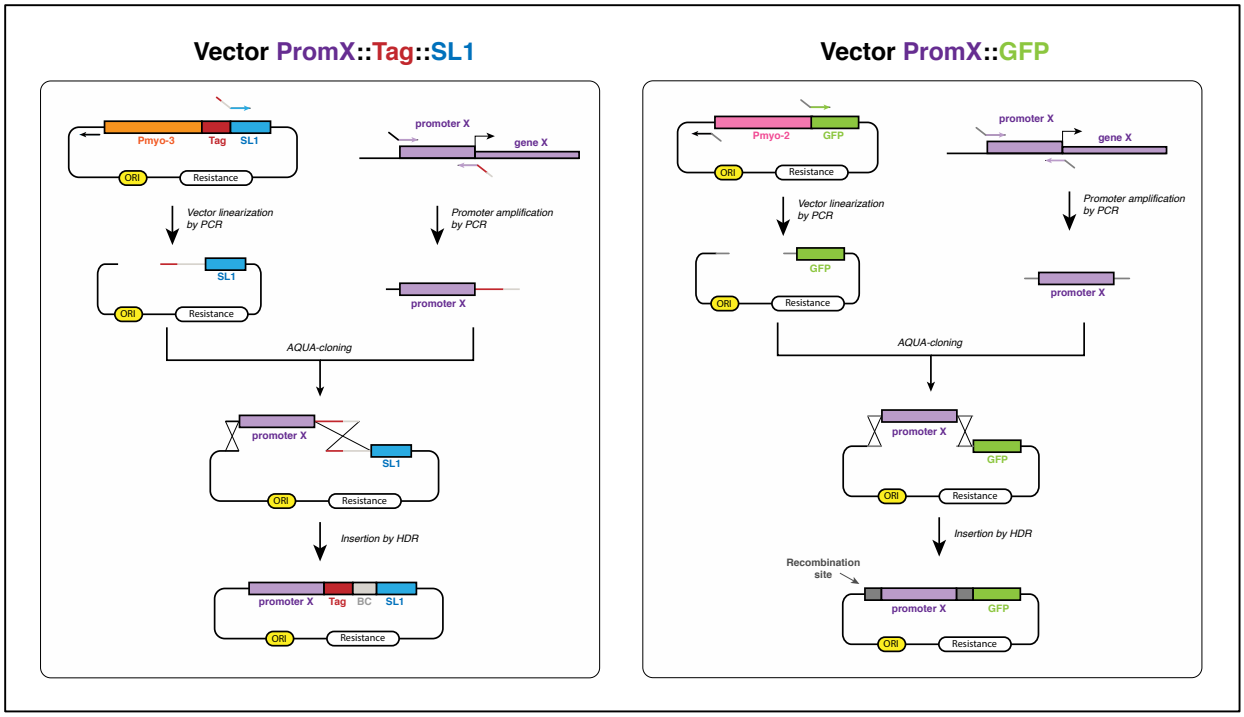


Figure 67 - AQUA-cloning strategies. The insertion of a tissue-specific promoter sequence in a Tag::SL1 construct (left panel) and in a GFP expressing construct (right panel).

An example of the validation of each step for the construction and verification of the plasmid Pgcy-7::Tag::SL1 is shown in **Figure 68**.

The amplification of the promoter sequence and the linearization of the vector is shown in **panel A**. The screening of 5 different clones after transformation was done by colony PCR (**panel B**) using M13 primers that are flanking the whole construct (Promoter::Tag::SL1). From this we obtained two clones for which the size of the amplification corresponded to the expected size of 1.9Kb for the insertion of Pgcy-7. After plasmid extraction (QIAprep Spin Miniprep Kit), we carried a double digestion using the restriction enzymes BsaI and SpeI (**panel C**). The backbone of the vector - common to all constructions - contains a BsaI restriction site, but the SpeI site is only carried by myo-3 promoter, allowing us to easily discriminate between the two constructs (**Figure 68.C**)

The same protocol was used for the construction and validation of the different vectors. For each step, primers and restrictions enzymes were selected in function of the different promoter sequence. The main 6 steps for the construction and validation of each vector construct is summarized in **Figure 69**.

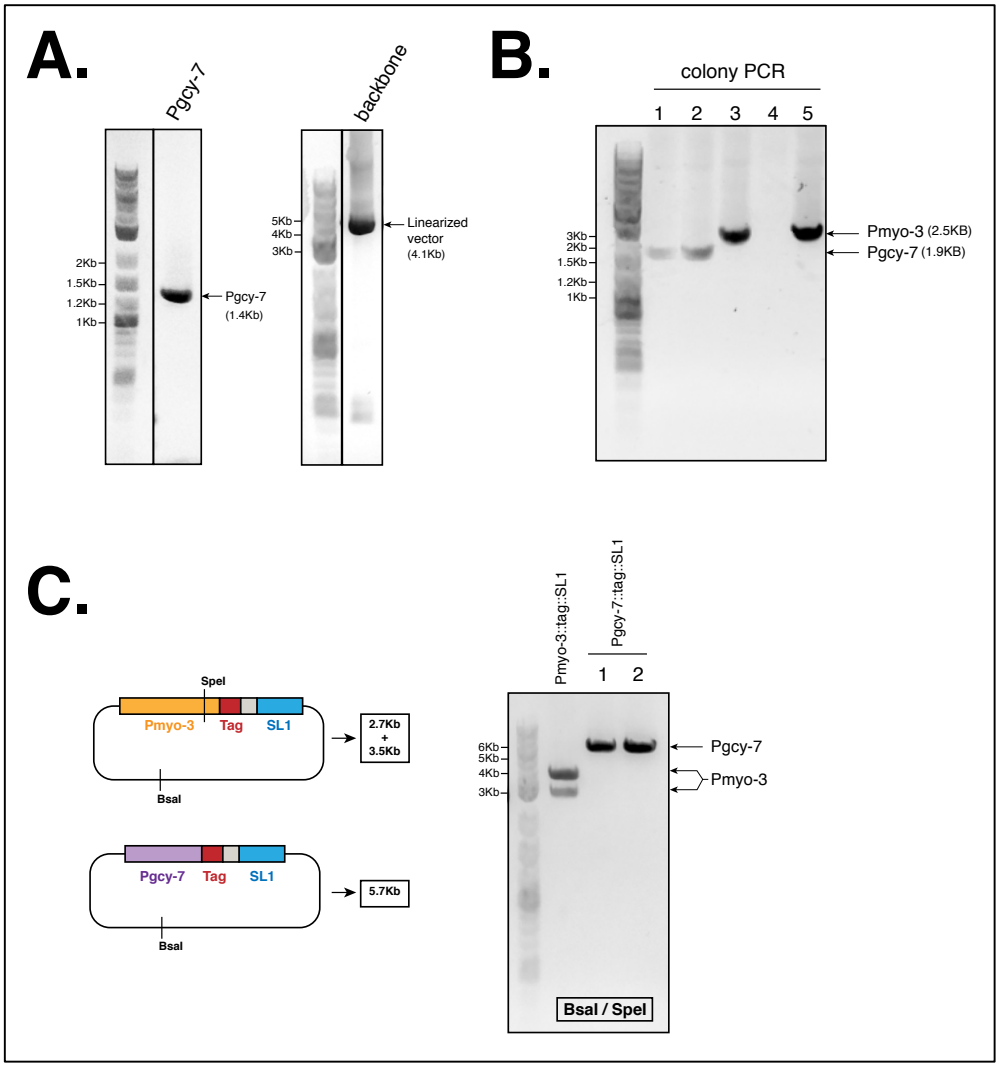


Figure 68 - Construction of Pgcy-7::Tag::SL1 vector. A) Amplification of Pgcy-7 and linearization of the vector backbone by PCR. **B)** Screening of different clones by colony PCR using M13 primers. **C)** Validation of the insertion of Pgcy-7 by double digestion with BsaI and SpeI restriction enzymes.

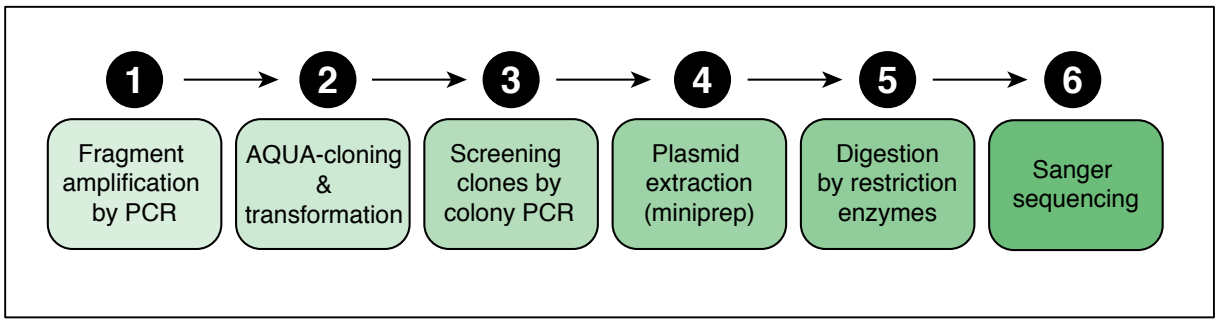


Figure 69 - Main steps for the construction and validation of a new vector construct

2.3.3 - Validation of the different promoters

While we finished to validate the different insertions, we started to generate extrachromosomal strains by performing micro-injections in WT worms. We started to inject constructions for which we had both set of vectors (Tag::SL1 and GFP expression), along with a plasmid carrying a G418 resistance. This resistance allows to use NGM+G418 plates in order to screen for worms carrying the extrachromosomal array. Those worms will develop on such plates but WT worms will not survive. This method allows for a rapid screening of the progeny when no GFP fluorescence is easily visible under the binocular (small tissues requiring the use of a microscope) but also ensures us to work with a population mostly composed of transgenic worms before proceeding with RNA extractions irrespective of the transmission rate of the extrachromosomal array.

We obtained three transgenic strains expressing the modified SL for which we could validate the expression of the promoter (**Figure 70**):

- Body wall muscles expression (promoter myo-3 from the initial construct)
- Pharyngeal expression (promoter myo-2)
- Pan-neuronal expression (promoter rgef-1)

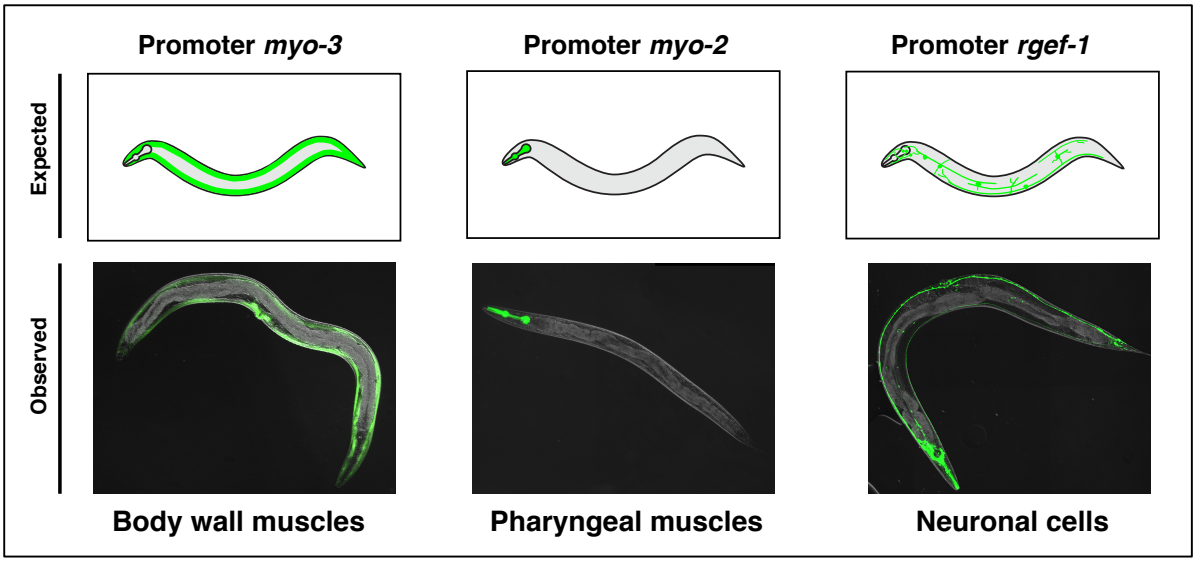


Figure 70 - In vivo activity of three different transcriptional reporters expressing GFP under the control of various tissue-specific promoter. Top panel represents the expected GFP expression and bottom panel represents the observed activity *in vivo*.

2.3.2 - Identification of tissue-specific RNAs

a) Amplification and isolation of Tagged RNAs from total RNAs extracts

After we generated and validated the first extrachromosomal strains, we let worms to grow on large NGM plates supplemented with G418 for 3 days. We verified under a binocular that most of the population were adults and expressing the fluorescent marker and then we harvested them for subsequent RNAs extractions. In parallel, we also performed RNA extraction from WT worms.

We used the total RNAs extracts from both strains (WT and Pmyo-3::Tag::SL1) for realizing different RT-PCR amplifications. First 1µg of each total RNAs extract was retro-transcribed using the Takara Primescript enzyme, for a total volume of 20µL. 2µL of each reaction was then used as template for PCR amplification. We amplified the *xbp-1* gene as a positive control and we then amplified all SL1 RNAs using a SL1 forward primer and an oligo-d(T) primer. A third amplification was performed on the cDNAs fraction coming from the Pmyo-3::Tag::SL1 strain. This amplification was done by using a Tag forward primer and an oligo-d(T) primer in order to amplify all of the Tagged SL1 RNAs present in the sample (**Figure 71.A**). The xbp-1 amplification was successful for both RNAs extracts, however we only noticed a faint SL1 smear and no Tag smear.

We first considered this might be due to a lack of material so we repeated the experiment - this time only on the RNAs extracted from the transgenic strain - by doubling the amount of template (4µL instead of 2µL previously). This time we could notice a clear SL1 smear, however we still could not observe a smear for the amplification of all Tag RNAs (**Figure 71.B**).

We then verified the lack of amplification for the Tag RNAs was not due to a problematic primer. To do so, we used a plasmid construct containing the Tag::SL1 construct and performed different amplifications:

- a SL1 amplification, using the SL1 forward primer and reverse primers
- a Tag::SL1 amplification by using the Tag forward primer and the SL1 reverse primer
- a positive control using M13 primers that allows for amplifying the backbone of the vector.

From this PCR amplification, we observed the correct amplification of all the regions, indicating the lack of amplification of Tag RNAs in previous RT-PCR experiments was not due to the Tag primer itself.

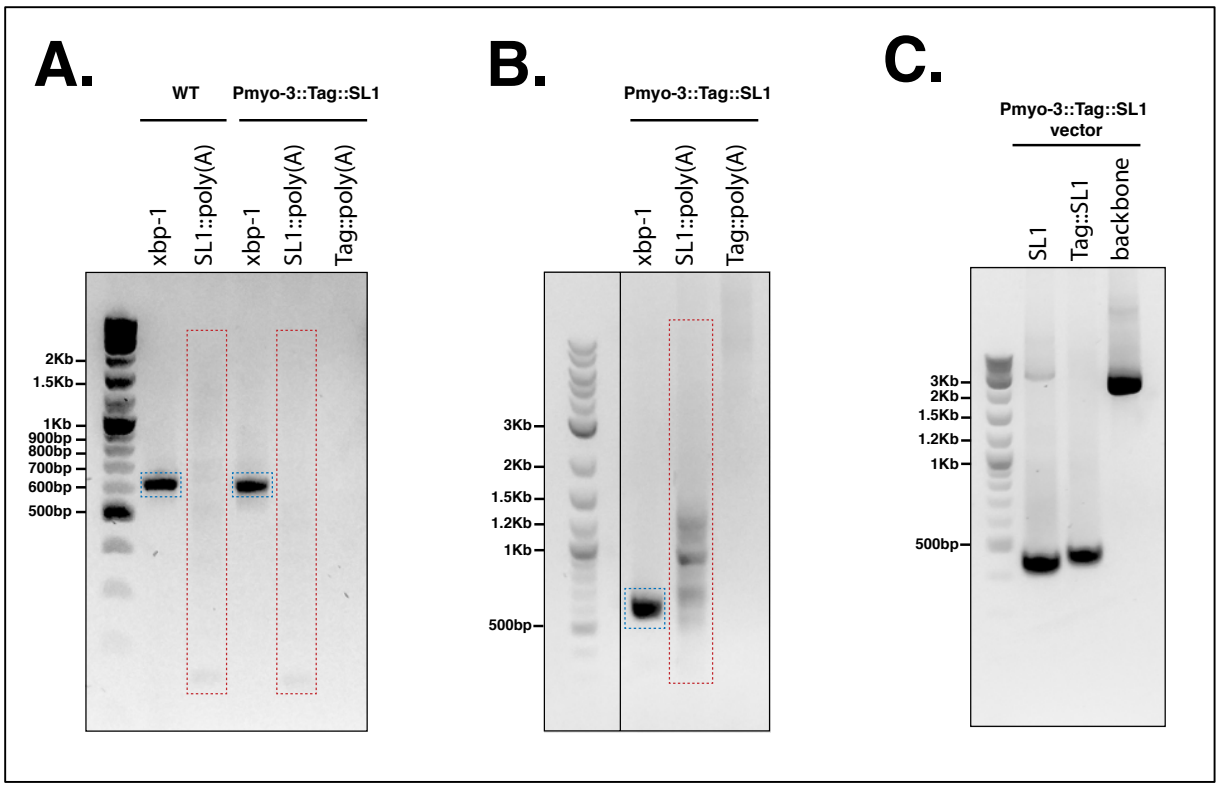


Figure 71 - Amplification by RT-PCR of SL1 and Tag RNAs. A) & B) RT-PCR for different amplifications (*xbp-1*, SL1 RNAs and Tag RNAs) on total RNAs extract from WT worms or transgenic corms carrying the Pmyo-3::Tag::SL1 extrachromosomal array. The red dotted square indicates the boundary of smears (faintly visible on gel). **C)** Validation of the Tag primer by amplifying a Tag::SL1 sequence from a verified plasmid.

Since we could not amplify Tag RNAs by RT-PCR, we decided to try to isolate those RNAs first as described by *Ma et al.* To do so, we used magnetics beads (MyOne Dynabeads C1) coated with streptavidin. The method works as follow: a biotinylated probe, complementary to a specific sequence (here the Tag sequence) is mixed with the magnetic beads and the RNAs extract. Following hybridization of the probe onto the target RNA, the biotin then interacts covalently with the streptavidin present at the surface of the beads (**Figure 72.A**). Then we use a magnet to pull all the beads on one side of the tube and the supernatant, containing RNAs not bound, is removed. By repeating the same procedure, the beads are washed several times in order to remove any non-specific hybridization. Finally, the captured RNAs are eluted in ddH_20 by denaturation with a treatment at 75°C for 2min.

Before performing the RNA pull-down, we verified the efficiency of the method. First, we tested the hybridization of a biotinylated Tag probe (called biotin-αTag) onto a Tag sequence or a SL1 sequence (as a negative control). In this experiment, we used Tag and SL1 primers on which were added a cyanine dye (Cy3). This dye can be excited at 532nm by using a Typhoon Trio+ which allows to detect smaller concentrations of primers than it would be otherwise possible with a treatment using ethidium bromide and UV exposure.

500nM of Cy3-primer (SL1 or Tag) were incubated with either 0, 250 or 500nM or biotin-αTag for 4h at 55°C. Each mix was then deposited on a 3% agarose gel (**Figure 72.B**). As in a gel shift assay, probes that hybridized on their target will run slower inside the agarose compared to probes that did not hybridize, allowing us to discriminate the conditions for which the hybridization was successful.

As expected, when using the Cy3-SL1 template and the biotin-αTag we do not see any hybridization. When using the Cy3-Tag template at 250nM with 500nM or probe we can see two bands, suggesting that all of the probe (250nM) was hybridized and the rest (250nM) was not, due to a lack of available template. When using the probe at 500nM, only a single band is visible, indicating that all of the probe was hybridized. From this assay, we could confirm our ability to capture the target of interest.

The second test we performed was related to the binding of the biotinylated probe onto magnetic beads coated with streptavidin. 6µL of Dynabeads MyOne C1 were washed three times in 500µL of 1X Washing Buffer (WB) and then re-suspended in 30µL of 2X WB. The beads were finally split in three different tubes and each tube was added with 10µL of a mix probe/primer from the previous experiment and incubated for 30min at room temperature under constant agitation (**Figure 72.C**):

- 500nM Cy3-SL1 with 500nM of biotin-αTag (0% hybridization)
- 500nM Cy3-Tag with 250nM of biotin-αTag (50% hybridization)
- 500nM Cy3-Tag with 500nM of biotin-αTag (100% hybridization)

After incubation, we pulled-down the beads with a magnet and retrieved the supernatant. For each condition, we deposited onto an agarose gel the mix pre-incubation (input) and the supernatant retrieved after incubation (output) and compared both fractions. In the first tube (0% hybridization), all of the template Cy3-SL1 was found in the supernatant, indicating it was not pulled-down along with the beads due to their lack of hybridization with the biotinylated probe. In the two other conditions (50% and 100% hybridization, respectively), only the non-hybridized template was found in the output, which confirmed that only the template hybridized with the probe was specifically pulled-down with the magnetic beads.

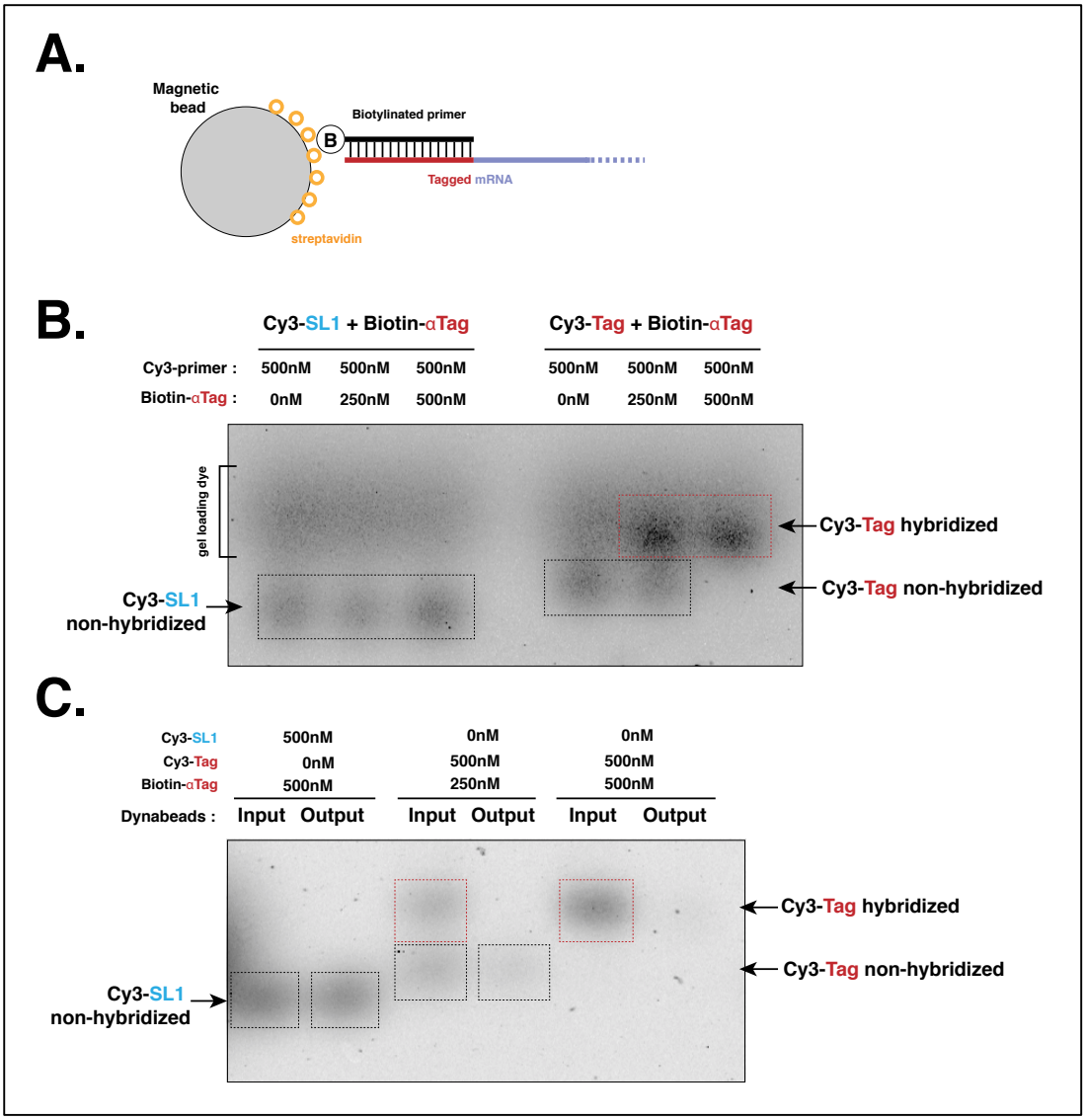


Figure 72 - Test of hybridization between Tag sequence and a biotinylated αTag sequence. **A)** Principle of isolations using a biotinylated probe and magnetic beads. **B)** Gel shift assay using a Cy3-primer (SL1 or Tag) and a biotinylated-αTag probe. **C)** Dynabeads supernatant (output) deposited on gel and compared with the input product (non-hybridized probes versus hybridized probes).

Following those initial tests, we tried to optimize the protocol for working with RNAs without risking degradation since the hybridization step was initially performed at 55°C for 4h. We decided to use a hybridization buffer (SSC buffer) to reduce both the incubation time and the temperature at which it is incubated. This time, we performed the hybridization at room temperature and tested three different incubation time (10min, 20min and 30min) and two concentrations of Cy3-Tag and Biotin-αTag (1µM and 10µM each). The probe and target were mixed in equimolar ratio (1:1) and a control was performed by omitting the addition of the probe (1:0 ratio).

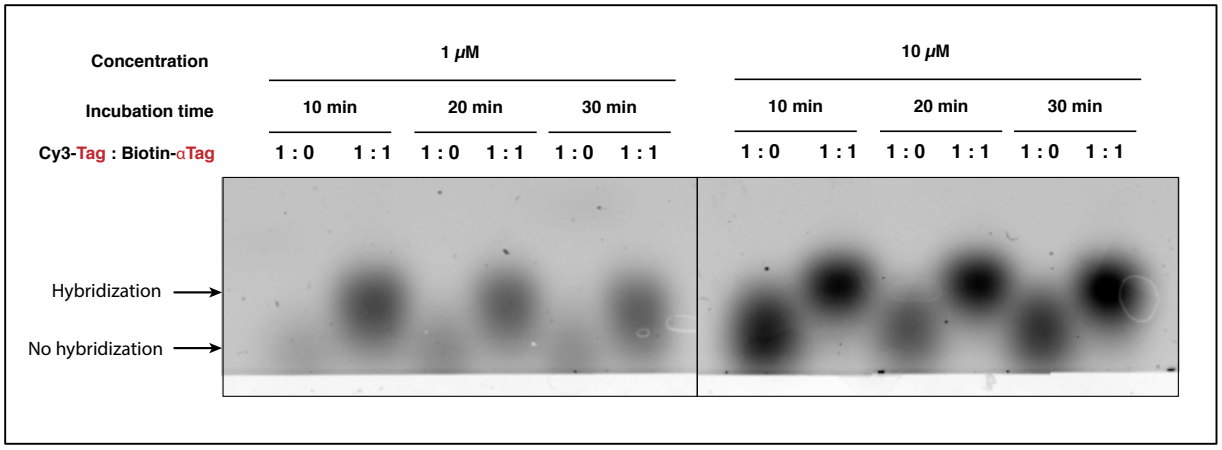


Figure 73 - Hybridization of Cy3-Tag and Biotin-αTag in SSC buffer. Two concentrations (1μM and 10μM) and three different incubation times (10min, 20min and 30min) are tested.

As done previously, each mix was run on a 3% agarose gel and Cy3-primers revealed on a Typhoon Trio+ at 532nm wave length. From the gel, we can observe that the hybridization was successful for all three incubation times tested and both concentrations (**Figure 73**), suggesting we could perform the hybridization step in 10min at room temperature, allowing us to work in conditions less prone to conduct to RNA degradation.

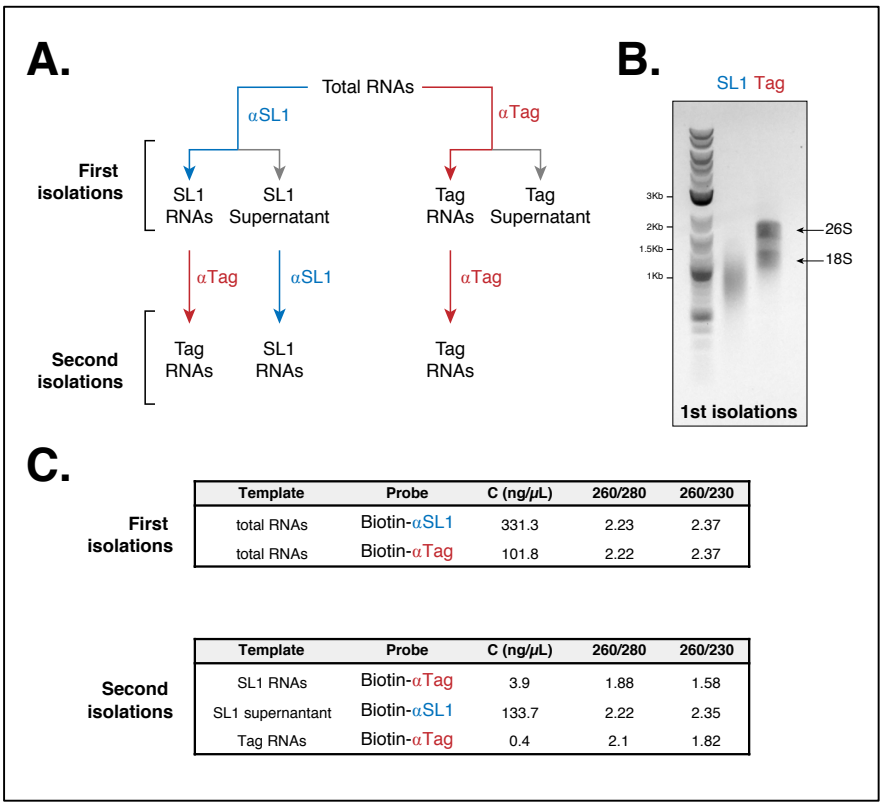
Based on those results, we tried to isolate SL1 and Tag RNAs. This time, we used total RNAs extracted from a strain expressing Pmec-7::Tag::SL1. Since this promoter is active in a lower number of cells (6 touch neurons) compared to the *myo-3* promoter (95 cells), we decided to perform the isolations on a large sample of total RNAs (200μg) in order to compensate for a lower number of Tag RNAs. Isolation were performed with either the biotin-αSL1 probe (SL1 isolations) or the biotin-αTag probe (Tag isolations) (**Figure 74.A**). After incubation of the RNAs, the supernatant of each reaction was retrieved for later use. Beads were then washed three times and eluted in 25μL of 10mM Tris-HCl.

The RNAs were quantified by spectrophotometry (**Figure 74.C - first isolations**) and 1μg of each eluate was deposited on an agarose gel for verifying their quality (**Figure 74.B**). From this, we observed a smear for the SL1 fraction similar to the smear observed in RT-PCR experiments, suggesting we managed to isolate SL1 RNAs. However, in the Tag fraction, we mainly observed two bands corresponding to the 16S and 28S ribosomal RNAs, indicating a non-specific isolation.

Based on those observations, we decided to perform a second round of isolations (**Figure 74.A**):

- A Tag isolation from the enriched SL1 fraction.
- A Tag isolation from the Tag RNAs, in order to get rid of contaminants rRNAs.
- A SL1 isolation from the "SL1 supernatant", in order to see if we had captured all SL1 RNAs in the first round of isolations.

The second round of isolation was performed in the same conditions as for the first round and once the RNAs were eluted, we quantified them by spectrophotometry (**Figure 74.C - second isolations**). From this quantification step, we determined that we were not able to isolate Tag RNAs since the concentrations were very low. Furthermore, both 260/280 and 260/230 ratios were out of the expected range (expected: 260/280 = 2 and 260/230 = 1.8-2.2).
We also observed that SL1 RNAs had been isolated from the supernatant, indicating that the first SL1 isolation had not been sufficient to capture all of the SL1 RNAs.



C.

	Template	Probe	C (ng/µL)	260/280	260/230
First isolations	total RNAs	Biotin-αSL1	331.3	2.23	2.37
	total RNAs	Biotin-αTag	101.8	2.22	2.37

	Template	Probe	C (ng/µL)	260/280	260/230
Second isolations	SL1 RNAs	Biotin-αTag	3.9	1.88	1.58
	SL1 supernantant	Biotin-αSL1	133.7	2.22	2.35
	Tag RNAs	Biotin-αTag	0.4	2.1	1.82

Figure 74 - Tag and SL1 RNAs isolation using magnetic beads. A) Description of the different isolations performed in the first and second round. **B)** SL1 and Tag isolations deposited on 2% agarose gel for quality control. **C)** Quantification of isolated RNAs by spectrophotometry. Blank was performed with the same buffer in which the RNAs are eluted.

Since we could not isolate Tag RNAs directly, we then tried to amplify them by RT-PCR from an enriched SL1 RNAs fraction. We tried to amplify a ubiquitously expressed gene (*rpl-5*) with both the SL and Tag primers, and we tried to amplify all of SL1 RNAs and all of Tag RNAs in the sample, as previously done. We managed to obtain an amplification with the SL1 primer but not with the Tag primer (**Figure 75**).
Our inability to amplify or isolate Tag RNAs seems to indicate that either the expression of the transgene is too low to be detected - a strong possibility considering we are working with the *mec-7* promoter which is expressed in a small set of cells. Yet, the lack of amplification by RT-PCR from RNAs extracts coming from the Pmyo-3::Tag::SL1 strain also suggests that modified SL1 are not trans-spliced onto mRNAs.

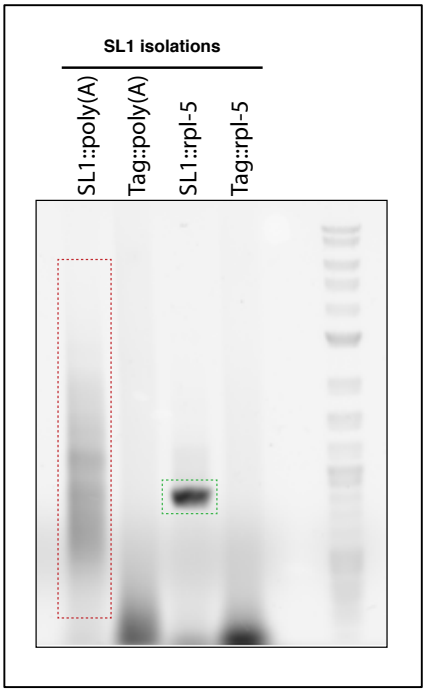


Figure 75 - Amplification of SL1 and Tag RNAs from an SL1 enriched fraction. Successful amplification is highlighted in red (SL1 smear) and in green (*rpl-5*).

b) Sequencing of a Pmyo-3::tag::SL1 strain

Since we were performing direct-cDNA sequencing experiments (see **section 2.1.2**), we decided to sequence the strain expressing the Pmyo-3::Tag::SL1 transgene to see if we could detect reads with a modified SL1 sequence.

The Tag sequence is expected to be found directly upstream the SL1 sequence so I used the same method than for finding SL reads. First I extracted the last 100bp of the 5' soft-clip sequence from every transcriptome-aligned read and performed a semi-global alignment for both the Tag and the SL1 sequence. Since we were interested in detecting full Tag::SL1 sequence, I searched for the full sequence and only considered reads which exhibited a high score (superior to 15) for both sequences. However, with those criteria, I could not find transcriptomic alignments containing a Tag::SL sequence (**Figure 76.A**).

Next, I decided to look at reads aligning onto the GFP. Since the GFP is expressed under the control of the same promoter than the Tag::SL1 construct, we hypothesized its mRNA might be more likely to be found associated with a Tag since they both are driven by the same promoter. I could only find a few hundred reads mapping to the GFP sequence, however when searching for the Tag::SL1, we only managed to detect 2 reads (from two independent experiments) (**Figure 76.A**) and upon visual inspection of the sequences I was able to confirm they were indeed Tag::SL1 reads (**Figure 76.B**). This observation confirms that our construct is able to be expressed and to

be found associated with other mRNAs, however due to the extremely low number of reads we could confirmed with this approach, it seems to indicate that a very small proportion of reads are actually trans-spliced with the modified SL1 RNA which could explain why we were not able to isolate or even amplify a library of Tag RNAs.

A.

		Transcriptome		GFP	
Experiment	**#**	**total reads**	**Tag reads**	**total reads**	**Tag reads**
SSP [polyA]	**1**	856 699	0	901	0
	2	265 555	0	177	0
	3	215 150	0	211	1
SSP [sl1]	**1**	329 046	0	36	0
	2	75 874	0	45	0
	3	520 347	0	119	1

B.

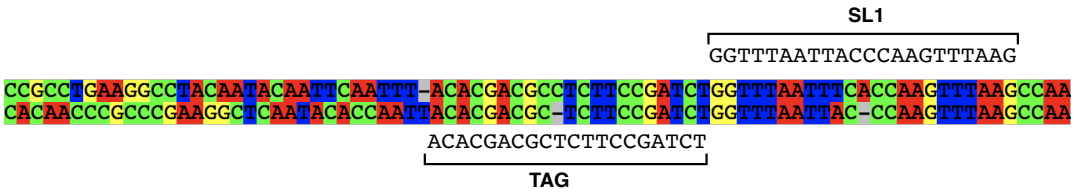


Figure 76 - Searching Tag::SL1 RNAs in sequencing reads. A) Number of reads for which we could detect a Tag::SL1 sequence for six different sequencing experiments. **B)** GFP reads for which we could detect the Tag::SL1sequence.

Chapter 3

Discussion

3.1 - The interest of using nanopore technology for transcriptomics

Nowadays, most methods for performing transcriptomics analysis still heavily relies on the use of short reads (Illumina sequencing). During my thesis, I aimed at using nanopore technology for generating long reads that would allow to directly address the question of combinatorial use of alternatively spliced exon in genes with multiple alternative splicing events, we also wanted to evaluate if avoiding PCR amplification could provide a more accurate picture of the reality of the cells.

a) Why choosing direct-cDNA sequencing ?

Nanopore sequencing being a very recent technology, we could not rely on pre-established protocols for performing nanopore sequencing in *C. elegans* and, therefore, we had to perform our own comparative analysis for determining which kit was more adapted to our needs.

In this study, we chose to test all three kits offered for RNA sequencing: a PCR-based kit (1D ligation products) and two PCR-free kits (direct-RNA and direct-cDNA).

It was interesting to test this kit since it is one of the first kit released by ONT and still is the most commonly used for genomic sequencing. In that sense, it allowed us to test the full extent of the technology and to directly compare with the other kits. On the other hand, if direct-RNA sequencing seemed an approach perfectly fitted for the study of gene expression, our results finally encouraged us to perform direct-cDNA sequencing due to the lower yield of the protocol and lower quality of reads obtained.

We have concluded the direct-cDNA kit was the performing best in term of quality of reads, sequencing output and conservation of genes ratios. It offered the reliability of using DNA-based kit (faster sequencing speed leading to higher throughput and better algorithms for basecalling leading to better quality sequences) while still conserving the advantage of RNA-based sequencing kit (absence of PCR bias, straightforward protocol).

b) Future prospects for direct-RNA sequencing

However, those results need to be put back in the context of the state of nanopore technology four years ago. At the time, nanopore sequencing had just released their first kit for performing direct-RNA sequencing and was still in the process of mastering new chemistry and new basecalling model for RNA sequencing. This lead us to observe sub-par results when compared to DNA-based kit but, given the speed at which technology improves, it is likely this kit has already been greatly improved over the last years.

Furthermore, in the context of the global pandemic of coronavirus SARS-Cov-2, nanopore has put an accent on the development of direct-RNA sequencing for the rapid identification of viral genomes and, those kits, might soon provide results equivalent - if not superior - to DNA based kits.

In the coming years, it is likely direct-RNA sequencing will become the new standard for the study of transcriptomics, particularly in *C. elegans*, where we observed artefacts directly linked to library preparation for direct-cDNA sequencing.

c) Direct-cDNA reads of *C. elegans* present library artefacts

We are reporting that the presence of a SL1 sequence on the 5' extremity of the messenger RNAs is able to generate a hairpin linking the two strands of cDNAs during library preparation. This hairpin then blocks the ligation of sequencing adapters on the 5' extremity of the ds-cDNA, which leads to a strong strand bias in which the antisense strand is preferentially sequenced. Moreover, this hairpin also contributes to the generation of abnormally long sequences, in which the sense strand of the cDNA is found to be directly linked to the 5' extremity of the antisense strand, resulting in long 5' soft-clip sequences.

This artefact does not affect our ability to identify a messenger RNA and the potential alternative splicing events within its sequence, however it hinders our ability to correctly detect splice-leader sequences on the 5' extremity of *C. elegans* mRNAs.

Nonetheless, we are reporting that the prevalence of this hairpin structure can be reduced in a given dataset by swapping the SSP primer used for universal 2nd strand synthesis in ONT protocol by a SL1 primer. By hybridizing on its complementary sequence on the first strand of cDNA, the SL1 primer allows to destabilize the secondary structure of the SL hairpin and makes the 5' extremity available for ligation of sequencing adapters. If this approach did not completely removed reads containing hairpins, it suggests that improved conditions (higher concentration of SL1 primer, increased temperature, etc.) for the binding of the SL1 primer on its target might be sufficient to do so if one desires, although it must be noted that generating libraries without strand bias would not necessarily provide more information content. Each messenger could be read twice (once on each strand) but since we are not sequencing all the molecules from the sample the distribution of gene frequencies would not be affected.

3.2 - Studying trans-splicing events

a) Is trans-splicing in *C. elegans* a ubiquitous mechanism ?

With this study, we report findings that correlates with previous observations related to the pervasiveness of trans-splicing in *C. elegans* (Tourasse et al, 2017).

Our analysis has allowed us to show trans-splicing activity in several genes, including genes that had not been previously found associated with a SL sequence. Interestingly, genes for which we could not detect trans-splicing activity exhibit a very low level of expression in our dataset.
This suggests that trans-splicing is a more ubiquitous mechanism than previously thought and that our ability to detect such events is only limited by the sensitivity of the methods used, thus bringing another argument in favour of the generation of tissue-specific transcriptome maps and single-cell analysis.

The study of gene expression in smaller tissues might allow to detect more easily the rare isoforms that are specific of their cell type and give us the possibility to uncover new genes subjected to this biological process.

b) The case of *lev-11* short and long isoforms

The study of *lev-11* isoforms has allowed us to uncover a surprising mechanism where two alternative promoters leads to the generation of differentially trans-spliced populations of mRNAs. If we do not know yet what is the exact functions of the different trans-splicing sequences in *C. elegans*, this observation provides another evidence in favour of a physiological role for those sequences.

Furthermore, the presence of hairpins at the end of all the mRNAs, even in the long isoforms of *lev-11* in which we could not detect any SL1 or SL2 sequences, raises some interesting questions regarding the ability of SL sequences to forms hairpins.

We could not clearly determine if the 18nt sequence found on the 5' extremity of *lev-11* long isoforms are a new SL sequence that is specifically found on those isoforms, or if these transcripts possess the innate ability to form a hairpin on their own. However, it is possible the 5'UTR sequence of *lev-11* long form acts as a biological mimic of the SL sequence which gives the ability of these transcripts to escape trans-splicing.
In such case, the study of SL-hairpins might bring new evidences regarding the physiological importance of trans-splicing.

3.3 - The interest of performing tissue-specific transcriptome analysis

During my thesis, I attempted at producing a tissue-specific transcriptome map for several tissues of interest of *C. elegans*. This work was initially motivated by the publication of a method supposedly allowing for tagging and isolating RNAs from any tissue of interest through the hacking of the trans-splicing machinery and the use of tissue-specific promoters (Ma *et al, 2017*).

The advantage of this method, over others methods like single-cell RNA-seq or the spatial transcriptomics approach developed by Ebbing et al, lies in its experimental protocol that is less complex and much more affordable. Indeed, this method only requires the generation of transgenic strains carrying extrachromosomal arrays - a protocol well established in the lab - and benefits from the large collection of promoter sequences that have been extensively characterized over the last 20 years. Furthermore, the versatility of this system makes it possible to study new tissues of interests as needs arise.

Initially, the authors considering that trans-splicing was targeting only about 85% of all *C. elegans* genes, thought that their approach could have fail to detect a significant portion of the transcriptome and provide providing an only a partial snapshot of *C. elegans* transcriptome. However, with our recent observations regarding the pervasiveness of the trans-splicing machinery, it is likely that all messengers are in fact targeted by trans-splicing. Therefore, we thought the method could be worth pursuing.

We have decided to combine this method with the new possibilities offered by nanopore technology to further study gene expression and uncover tissue-specificity between different cellular types (muscles and neurons), however we were not able to correctly isolates the RNAs of interest. After sequencing of a strain expressing the modified SL1 in the body wall muscles, we were only able to detect two reads across six independent experiments with the Tag::SL1 sequence, confirming that our inability to amplify or isolate those RNAs was due to the extremely low level of RNAs presenting the Tag sequence.

If our result tends to confirm it is possible to hack the trans-splicing machinery in order to modify RNAs presented in a given cell type, one can wonder why the method did not work as efficiently for us compared to what is reported by Ma *et al.*
One possible explanation might the extensive use of PCR amplifications in their protocol and, especially, of nested PCR for increased sensitivity. However, due to the nature of our analysis, performing PCR amplification is not a desirable option and, therefore, other solutions will need to be found in order to adapt this protocol to direct-cDNA sequencing.
Despite using the same vector provided by the authors of the methods we generated our own transgenic line it is possible that the number of transgene copies in the extrachromosomal array is much lower in the line we selected compared to the one used in the original study.

Another difference between our protocol and theirs is the use of transcriptional reporters as co-injection markers. We expressed GFP under the control of the same promoter than the modified SL, which might participate in promoter titration, reducing the general level of expression of both constructs. Since those GFP-expressing constructs were primarily generated for validation of promoter activity, we could consider the generation of strains that only express the Tag::SL1 construct. However, when looking for GFP transcripts in our transgenic strains expressing both vector (Tag::SL and GFP constructs) there is not a strong level either, we only got 1489 reads mapping to GFP out of 6 different experiments (~2.1M reads mapping the transcriptome), among which we could only detect 2 reads which carry a tagged SL

It is likely that adapting this method for efficient direct-cDNA sequencing might require to remove the competition with the numerous copies of the endogenous SL1 gene (108 copies present on the genome).
This approach will be challenging due to the fact that SL1 has been showed to be essential for embryonic development, however is it possible to rescue the phenotype by expression of the SL1 gene as an extrachromosomal array. Therefore, we would need to specifically remove this gene in the tissue of interest and, instead, express our modified SL, then 100% of trans-spliced RNAs would incorporate the Tag sequence.

With the discovery of CRISPR/Cas9, a lot of progress has been made in the domain of genome engineering and such an approach is now conceivable. In *C. elegans*, protocols for conditional gene knockout have already been described (Shen et al. 2014). Their methods take advantage of the possibility to express exogenous DNA in the cells of the worm by injecting extrachromosomal arrays expressing the Cas9 protein and its guide RNA under the control of a tissue-specific promoter, which permit to carry out genome editing only in the tissue of interest. This approach might be heavily dependent on the amounts of cells in which genome editing will be successful, however for such a large tissue as body wall muscle, it is conceivable that successful events in a subpopulation of cells could be sufficient to isolate enough material to produce tissue-specific transcriptome maps.

However, before attempting such a complex approach, we can also consider to remove all but one copy of the endogenous SL1 gene. Since extrachromosomal arrays contains numerous copies of the transgene, this could allow us to revert the ratio of WT SL1 versus our modified SL1, hence allowing to tag more messengers.

3.4 - Future work: the study of exon associations

If this project was initially motivated by the study of exon associations within isoforms through the use of nanopore technology, due to the publication of the meta-analysis of exon junctions in *C. elegans*, much of my work was finally focused on the elucidation of the artefact observed in our sequencing dataset and the study of the trans-splicing mechanism.

To date, most of the available tools for the elucidation of alternative splicing events in nanopore reads relies on the clustering of similar sequences for the establishment of a set of uniquely represented sequences that are representatives of the different isoforms found in the dataset. However, such tools are not properly adapted to the study of exon's relationship.

During the last months of my thesis, I attempted to develop a new approach for finding isoforms and, particularly, for the study of exon associations. However, because of time constraints, no results could be presented in this manuscript. The study of exon association is of particular interest with nanopore transcriptomics datasets and might bring new insights regarding the regulation of alternative splicing events.

Chapter 4

Material and methods

4.1 - Solutions and growth medium

4.1.1 - Solutions

1M Potassium Phosphate Buffer (KPO4)

9.5 g of KH2PO4, 5.25g of K2HPO4, ddH2O qsp 100mL. Adjust pH to 6.5.

S buffer:

2.925g of NaCl, 25mL of 1M KPO4 buffer and 475mL of ddH2O.

M9 solution:

3g of KH2PO4, 5.7g of NaHPO4, 5g of NaCl, 1mL of 1M MgSO4 and ddH2O qsp 1L. The obtained solution is filtered using a filtration unit.
This solution is used for manipulation and recovery of worms after micro-injection, egg-preparation, etc.

Liquid freezing solution:

The solution is made by diluting glycerol to 30% in S buffer. It is then autoclaved
This solution is used for storing worm strains at -80°C.

Worm lysis solution:

8.25mL of ddH2O, 3.75mL of 1M NaOH and 3mL of Bleach. (Do not use germicidal bleach).
This solution is used for egg-prep (« *bleaching* ») to synchronize a population of worms or remove contamination from a plate. It is prepared just before use and can be conserved up to one week at 4°C.

4.1.2 - Medium

Lysogeny Broth (LB):

10g of bactotryptone, 5g of yeast extract, 10g of NaCl and ddH2O qsp (*quantum statis*) 1L.
If solid medium is desired, add 7.5g of agar to the solution.
This medium is used for culture and maintenance of bacterial strains.
For antibiotic selection, 100µg/mL of Ampicillin or 50µg/mL of Kanamycin are added after autoclave or onto agar plates prior to use.

Super Optimal Broth (SOB):

10g of bactotryptone, 2.5g of yeast extract, 0.25g of NaCl, 0.09g of KCl and 5mL of 2M MgCl2. This medium is used for bacterial recovery after transformation and for allowing production of the resistance enzyme in case of selection with bactericide antibiotics.

Nematode Growth Medium (NGM):

3g of NaCl, 2.5g of peptone, 12g of agar, 1mL of cholesterol (5mg/mL in EtOh) and 975mL ddH2O. After autoclave, addition of 1mL of 1M CaCl2, 1mL of 1M MgSO4 and 25mL of 1M potassium phosphate.
This medium is commonly used for culture and maintenance of worms. Agar plates are "seeded" with a drop of solution of *Escherichia coli* OP50, as a source of food for *C. elegans*. The bacterial lawn is grown overnight before use.
For antibiotic selection, 0.4mg/mL of G418 (Geneticin) is added to the plate.

4.2 - Worm manipulations and transgenesis

4.2.1 - Handling worms

Maintenance and selection:

Worms are grown at 20°C onto NGM plates seeded with bacteria (Escherichia coli, OP50). In order to maintain the population, worms are transferred onto new plates every 3 days.

For selection, plates can be added with G418 and seeded with OP50 NeoR30. L2 larvae worms presenting the resistance will thrive on NGM+G418 plates while WT individuals will die, allowing for a quick and efficient selection method (Giordano-Santini and Dupuy 2011).

Freezing strains:

A large plate containing freshly starved L1-L2 worms is washed with M9 solution. Worms are recovered into a falcon tube and washed with M9 several times by performing gentle centrifugation (400g, 1min) and removing the supernatant. After washing, a volume of worms re-suspended in M9 is mixed into a cryo-tube with one volume of freezing solution and the tube are then stored at -80°C. After 24h, at least one vial is thawed to make sure the freezing procedure was successful and worms can be revived.

Worms strains that have been verified or brought from the Caenorhabditis Genetics Center (CGC) are kept at -80°C.

Egg-preparation: synchronization and decontamination of *C. elegans* strains:

Worms are grown 2-3 days on a seeded NGM plate in order to obtain gravid adults. 1-2mL of M9 solution is poured onto the plate to detach the worms from the agar, worms are then transferred to a 15mL falcon tube.

First, worms are washed. This step is performed by gentle centrifugation (400g, 1min) followed by removal of the supernatant without disturbing the worm pellet. 10mL of M9 solution is then added again and the step is repeated two more times.
After the final wash, 2mL of lysis buffer is added into the tube instead of M9 solution and the solution is mixed thoroughly. Lysis is carried out for a maximum of 5min and stopped if most of the worms can no longer be observed in the tube.
To slow down the reaction, 10mL of M9 solution is added to the lysis buffer. Then worms are quickly centrifuged (400g, 1min) and the supernatant is removed. 10mL of M9 solution is added

to the egg pellet and the previous step is repeated at least two more times to prevent residual lysis buffer to affect the eggs.

Eggs can then be distributed directly onto seeded NGM plates at 20°C or re-suspended in M9 solution overnight under gentle agitation. Due to the lack of food in M9 solution, eggs will hatch but larvae will arrest their development at L1 stage, resulting in a synchronized population of L1 worms. They can then be put onto seeded NGM plates where they will start to grow again.

Additionally, this protocol can be used to remove bacterial or yeast contamination from a nematode strain.

4.2.2 - Transgenesis and micro-injection

a) Trans-genesis strategy

- Extrachromosomal arrays:

10ng/µL of pRG02(NeoR) and 90ng/µL of transgene DNA (plasmid or PCR product).
If more than one transgene needs to be injected, they are mixed in equimolar concentration

- Micro-injection of worms

Worms are immobilized in a drop of oil onto a microscopy cover slip added with a 2% agarose pad. The injection is visualized under a Nikon Eclipse Ti microscope, linked to an Eppendorf Femtojet microinjector. The needle is loaded with 2µL of injection mix.

The injection takes place in the distal arm (syncytium) of one of the two gonads of *C. elegans*. After injection, P0 adults are recovered in M9 solution with a pipette and placed onto NGM plates (up to 5 for small ones, up to 10 for larger ones) with or without bacterial selection depending on the injection mix. Plates are put to 20°C and the progeny is observed under fluorescent light 2-3 days later.

Transgenic F1 individuals descending from P0 worms are isolated on separated plates supplemented with G418 for antibiotic selection.

4.2.3 - Microscopy

Before starting the observations, a 2% agarose pad is added onto a microscopy slide. Then a drop of 10mmol/L of levamisol is added onto the agarose. The worms are put into this solution in order to paralyze them without killing them and a cover slide is added on top of that. The observation is

made under Zeiss microscope Axio Imager Z1 equipped with a HXP 120 lamp and an Axiocam MRM camera for the acquisition of data. The worms are observed with optics 10x or 40x in bright field or DIC (Differential Interference Contrast Microscopy) and then under different fluorescent light (excitation at 395nm for GFP and 558nm for dsRed).

4.3 - Molecular biology

4.3.1 - Extraction of nucleic acids from *C. elegans* strains

Preparation of genomic DNA:

N2 worms growing on a NGM plate are recovered by washing the plate with 15mL of ddH_20 and transferred into a falcon tube. Worms are then washed by performing gentle centrifugation (400g, 1min) and removing the supernatant. ddH20 is added again onto the pellet to re-suspend worms and the previous step is repeated. A total of three washes is performed.
Then, genomic DNA extraction is performed using the NucleoSpin Tissue Kit from Macheray-Nagel. Worms are re-suspended in lysis buffer (T1) and frozen at -20°c for 1H. The sample is then thawed, 25µL of proteinase K (22.5 mg/mL) is added to the tube and the tube is incubated at 56°c overnight. The following steps are performed as indicated by the protocol provided by the manufacturer. The final elution is performed in 100µL ddH_20 pre-heated at 70°c.
The quality of the genomic DNA is evaluated on a Nanodrop machine by spectrophotometry and final concentration is adjusted with ddH_20 at ~50ng/µL.
Genomic DNA is then stored at 4°C.

Extraction of total RNAs:

a NGM plate containing gravid worms is washed with ddH_20 and worms are transferred into a falcon tube. Worms are then centrifuged gently (400g, 1min) and the supernatant is removed. Clean ddH_20 is added and worms are centrifuged again. This step is repeated at least one more time for a total of three washes. After the last centrifugation, the supernatant is removed and 1 volume of RNAble is added to the worm pellet and 1/10 volume of chloroform. At this point, the tube can be frozen at -80°c for later RNA extractions.

Otherwise, the tube is vigorously shaken for 1min and then let to rest on ice for 10min before centrifugation (14000 rpm, 15min at 4°c - this setting is used for the rest of the protocol).
The aqueous phase is then recovered and transferred into an Eppendorf tube. 1 volume of isopropanol is added and the tube is left at 4°c overnight for precipitation of nucleic acids. Another centrifugation is performed and the supernatant is removed without disturbing the pellet. A washing step is then performed by adding 1mL of 70% ethanol onto the pellet. After centrifugation, the supernatant is carefully removed and the pellet is left to dry at room temperature. Finally, the total RNAs are re-suspended in ddH_20.

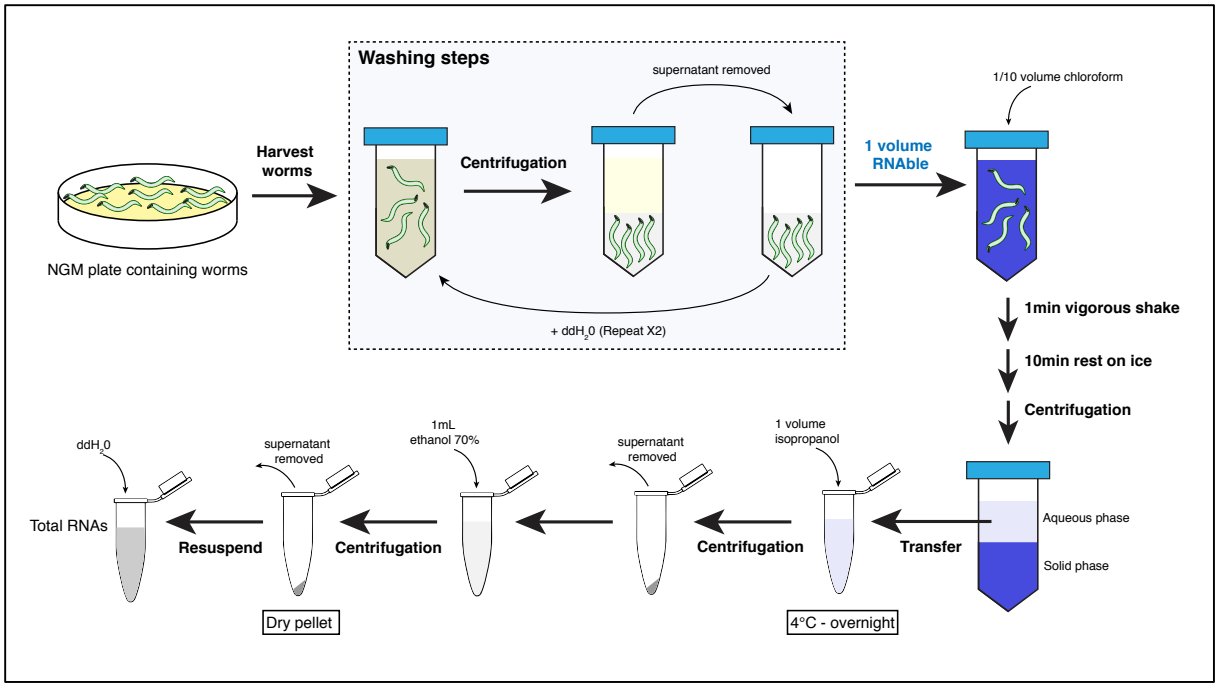


Figure 77 - Protocol for extracting total RNAs from *C. elegans* strains.

DNA/RNA quantification:

DNA (or RNA) samples are quantified by a spectrophotometer (NanoDrop). Before quantification, blank is performed with the buffer of the sample. 1µL of each sample is analysed by the machine. Samples are considered pure when the concentration is higher than 20ng/µL and if the ratios 260/280 and 260/230 are respectively comprised between 1.8-2.0 and 2.0-2.4 .

4.3.2 - Amplification of nucleic acids

a) Amplification

Amplification of DNA by PCR:

Unless clearly stated, PCR reaction are performed following the reaction mix described below in **Table 6**.

A.

Reagent	Volume
DNA template	1µL
Phusion buffer (5X)	10µL
dNTPs mix (10mM each)	1µL
Forward primer (10µM)	2.5µL
Reverse primer (10µM)	2.5µL
Phusion polymerase	0.5µL
ddH20	qsp 50µL

B.

Step	T°C	Time	Number of cycles
Initial denaturation	94°C	5min	1X
Denaturation	94°C	30s	30X
Hybridization	Tm	30s	
Elongation	72°C	30s/Kb	
Final extension	72°C	10min	1X

Table 6 - PCR amplification. A) Reaction mix **B)** Thermal cycler program.

RT-PCR:

cDNAs synthesis is performed using First strand cDNA synthesis kit from NEB. Add 1µg of total RNAs extract, 2µL of 50µM oligo-d(T)23-VN and dH20 qsp 8µL. Incubate 5min at 70°C. Add 10µL of M-MulV mix and 2µL of M-MulV enzyme. Incubate the 20µL of reaction at 42°C for 1H and then put it at 80°C for 5min in order to inactivate the enzymes. Negative controls are performed by replacing M-MulV enzyme with dH20. PCRs are then performed with 1µL of RT-PCR product (or negative controls) for 25µL of total reaction volume.

b) Control of amplification

Gel electrophoresis:

DNA samples are run on 1%-2% agarose gels (1g-2g of agar/100mL of TAE 0.5X) supplemented with Ethidium Bromide in order to visualize nucleic acids under UV lamp. 10µL of each sample is mixed with 2µL of 6X loading dye and loaded onto the gel, next to a molecular weight marker (1Kb DNA ladder or 2 log ladder).
The electrophoresis is performed in TAE 0.5X at 100V for about 40min.

Purification of PCR products:

Purification of PCR or RT-PCR reactions is performed onto purification columns (QIAGEN kit) following recommendations from the manufacturer.
Up to 1mg of DNA is loaded onto each purification column. PCR products are then washed with the provided buffers in order to eliminate the chemical compounds of the reaction (salts, proteins, etc.). DNA is eluted in 15-30µL of elution buffer and stored at -20°C.

4.3.3 - Generation of DNA plasmids

a) Plasmid assembly

Advanced Quick Assembly cloning (AQUA-cloning):

DNA fragments (inserts and linearized vector) are generated by PCR with 32bp flanking regions homologous to adjacent DNA fragments. They are mixed together in a final volume of 10µL ddH20 with molar ratios of 3:1 (insert : vector) and left at room temperature for 1h.
5µL of the reaction is used to transform 25µL of competent cells by heat shock (*E. coli*, DH5α).

Bacterial Transformation by heat-shock:

25µL of thermo-competent Escherichia coli cells (DH5α strain) is transformed with 5µL of DNA (PCR product, assembly fragments, etc.). Negative controls are performed with ddH20 instead of DNA. For heat shock, the mix of cells and DNA is kept on ice at 4°C for 30min, then brought up at 42°C for 1min and put back at 4°C for 5min. The bacteria are then put to recover at 37°C, under agitation, in 250µL of SOB for 30-60min. Finally, they are spread onto LB plates supplemented with proper antibiotic and left to grow overnight at 37°C.

b) Verification of plasmid constructs

Colony PCR:

This is used for screening colonies that integrated the plasmid construct.
Colonies are picked and transferred separately into 5µL of LB liquid medium. 1µL is used as DNA template for the PCR reaction and the rest is used for bacterial cultures.
The PCR reaction is carried out as previously described but the thermal cycler program is adjusted in order to ease the first round of amplification:

Step	T°C	Time	Number of cycles
Initial denaturation	94°C	5min	1X
Denaturation	94°C	1min	
Hybridization	Tm	1min	1X
Elongation	72°C	1min/Kb	
Denaturation	94°C	30s	
Hybridization	Tm	30s	30X
Elongation	72°C	30s/Kb	
Final extension	72°C	10min	1X

Table 7 - Thermal cycler program for colony PCR

Bacterial cultures:

After transformation, colonies that grew on selective plates are picked separately and put to grow in 5mL of LB + antibiotic at 37°C, with agitation, overnight.

Plasmid extraction (Miniprep):

The extraction of plasmids is performed using a QIAGEN Nucleo Spin Tissue kit.
2mL of each bacterial culture is centrifuged at 4000G for 10min. Supernatants are discarded and pellets re-suspended in 250µL of buffer P1. A lysis reaction is then started with the addition of 250µL of buffer N2. Tubes are mixed by inverting them 2-3 times. The reaction is then stopped by adding 350µL of buffer N3 and mixing the tubes by inversion. Finally, the tubes are centrifuged at 18000G for 10min and the supernatant is applied onto one the column provided with the kit.

The column is centrifuged for 1min at 18000G and the eluate is discarded. 500µL of PB buffer is added and the column is centrifuged again. 750µL of buffer PE is added and the column is centrifuged two times to make sure it is completely dry. Finally, the column is moved into a clean Eppendorf tube and DNA is eluted in 50µL of Elution Buffer.

Digestion by restriction enzymes:

~500ng of DNA is digested with 1U of enzyme in NEB buffer for at least 1H at the indicated temperature. Proper buffer and temperature of incubation is enzyme-dependent and must be verified for each reaction.
For double digest reactions, the NEB web-tool "Double Digest Finder" is used in order to determine the optimal conditions of the reaction (temperature, incubation time, etc.).

After incubation, each reaction is deposited into an agarose gel for electrophoresis. The obtained bands are compared with the expected profile of digestion. Plasmids that produced fragments of the expected length are sent to be sequenced for further validation.

Sanger sequencing:

DNA samples (plasmids or PCR products) are sent to be sequenced by Sanger reaction to GATC Biotech Company. In a single tube, 5µL of template DNA (80-100ng/µL for purified plasmids or 20-80ng/µL for PCR products) is mixed with 5µL of a primer at 5µM.
Each Sanger reaction generate between 800bp and 1000bp. Primers are chosen in order to sequence the construct entirely.

The chromatograms obtained are open using Snapgene viewer and the sequences are compared with the sequence of the expected construct.

4.3.4 - Nanopore sequencing

a) Library preparation

Preparation of sequencing libraries was performed according to the instructions provided by Oxford Nanopore Technology. An overview of each protocol - adapted from technical information found on ONT website - is provided for each of the three kits tested. Complementary information can be found at: https://store.nanoporetech.com

1D ligation kit:

For *C. elegans* libraries, retro-transcription and PCR amplification was performed independently from the library preparation described below. We used poly(T) and SL1 specific primers for RT-PCR of all SL1-mRNAs.

The end of the protocol was then strictly followed for prepping cDNA extremities and performing ligation of the sequencing adapters.

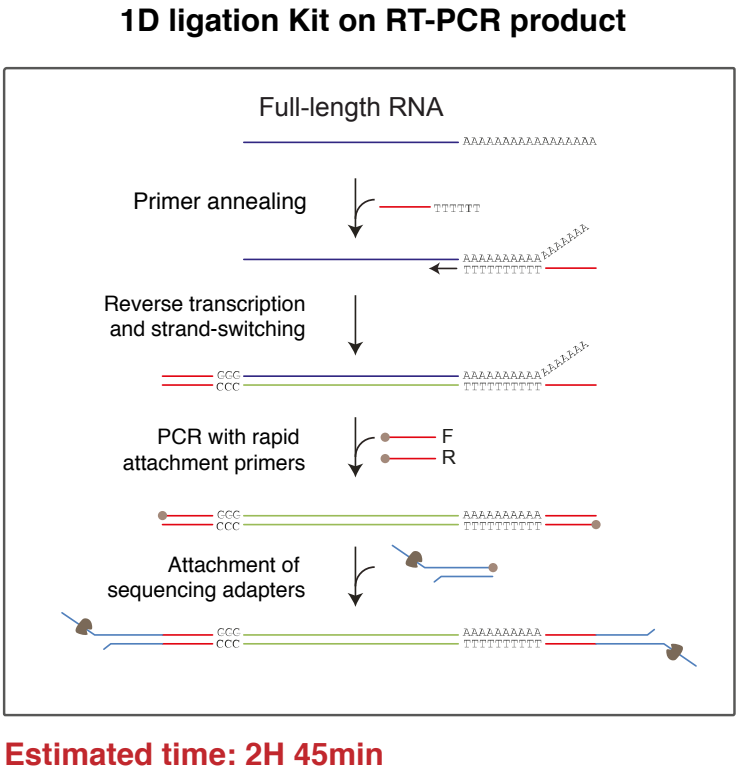


Figure 78 - Schematic overview of library preparation for 1D ligation kit (Adapted from ONT website).

Direct-cDNA sequencing kit:

Unless clearly stated, the protocol as sketched below was strictly followed for library preparation.

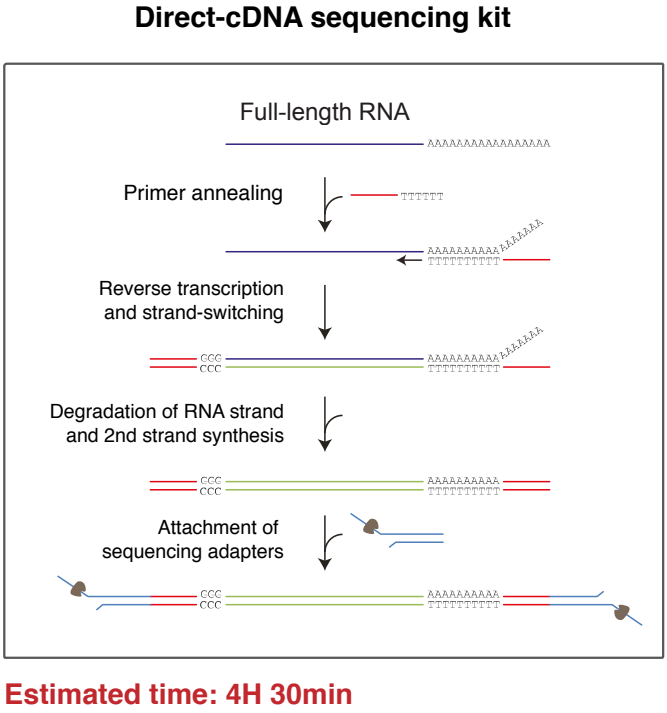


Figure 79 - Schematic overview of library preparation for Direct-cDNA sequencing kit (Adapted from ONT website).

Direct-RNA sequencing kit:

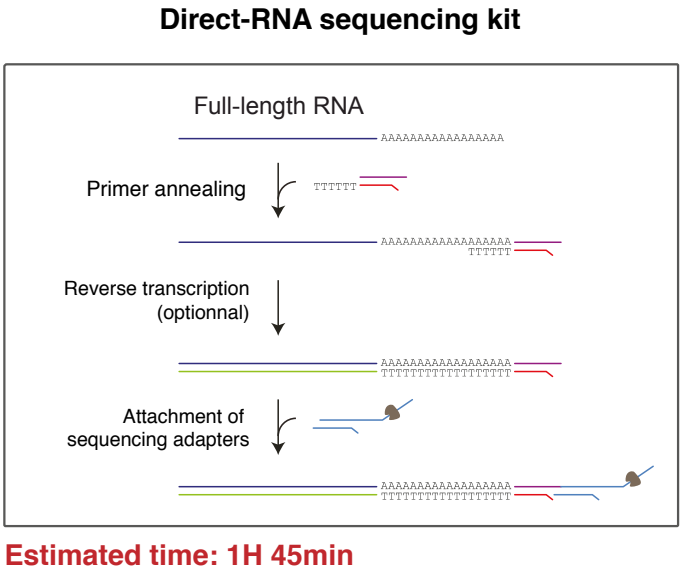


Figure 80 - Schematic overview of library preparation for Direct-RNA sequencing kit (Adapted from ONT website).

b) Preparation of MinION device and flowcell

All of the samples were performed with a MinION device (Mk1B). Samples were loaded onto FLO-MIN 106 flowcell (pores type: R9.4) bought from ONT. The device was plugged onto a computer via USB and controlled through the MinKNOW software. All sequencing experiments were generated on site.

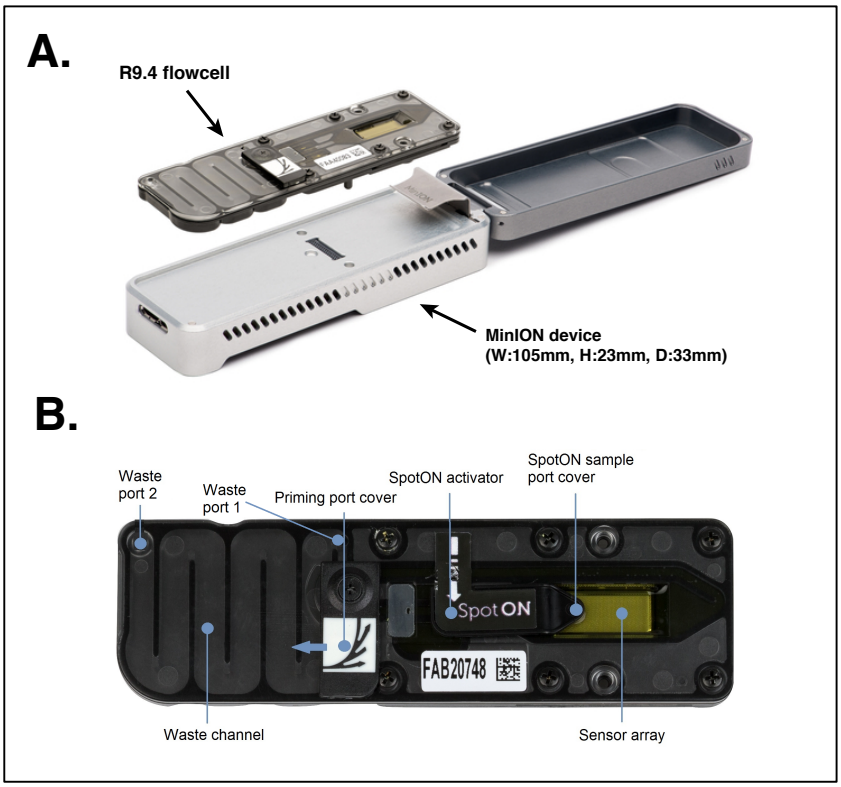


Figure 81 - MinION sequencer and flowcell chip. A) MinION device and flowcell. **B)** R9.4 flowcell and the different ports.

Quality control of flowcell:

When using new flowcells, a quality control procedure must be performed prior to sequencing. Flowcells are pre-loaded with a QC DNA molecule present in its storage buffer. The molecule will produce a distinctive nanopore signal that will be recognized by the MinKNOW software for validating the integrity of each nanopore before use. At the end of the procedure, the number of active pores is reported. If the number is too low, the flowcell can be returned to the manufacturer for replacement. This step ensure that sequencing is performed in the best conditions possible.

Flowcell priming:

This step is performed prior to library loading. It allows to replace the storage buffer with a priming buffer made by mixing the Flush buffer (FB) with the Flush tether (FT) provided in the kit. During this step, it is essential to avoid the introduction of air bubbles as this would damage the integrity of the pores.
The exact procedure for priming the flowcell is detailed in every sequencing protocol and on ONT website.

Library loading:

Following library preparation, a pre-sequencing mix is made up. This mix contains the library, water, Sequencing Buffer (SQB) and Loading Beads (LB). It is then loaded into the sample port of the flowcell, drop by drop. Once the pre-sequencing mix is fully loaded, the sample port is closed and the sequencing experiment can be started through the MinKNOW interface.

Running the sequencing experiment:

The MinKNOW interface allow the user to configure the experiment based on the library being sequenced: type of kit being used, length of the run, activation of real-time basecalling, output files, etc.

Once the experiment is started, different statistics are shown in real-time, such as the status of each of the channel on the flowcell (sequencing, available, inactive, etc.), the number of reads sequenced and their length. This allows to monitor the quality of the run and abort the experiment if there is any problem (lots of inactive pores, poor library, etc.).

Flowcell washing and storing:

Multiple libraries can be run on the same flowcell. Therefore, after a sequencing run, the remaining library needs to be flushed out. This step allows to remove most of the previous libraries (99.9% according to ONT documentation). After washing of the flowcell, it can be directly reused by loading a new library or can be stored at 4°C for later use. In case of storing, a Storage Buffer is introduced into the flowcell before for keeping the integrity of the nanopore array while stored.

4.4 - Bioinformatics

4.4.1 - Pre-processing data

a) Basecalling

After acquisition of the raw data (FAST5 files), basecalling was performed using the guppy basecaller provided by ONT. The following command line was used:

```
guppy_basecaller -i path/to/fast5/directory \
-r -s path/to/output/directory -c dna_r9.4.1_450bps_fast.cfg \
--trim_strategy none --qscore_filtering 1 --calib_detect 1 -x auto
```

Description of the arguments:

-i : path to fast5 files
-r : allow to search files in subdirectories
-s : output path for the fastq file.
-c : model used for basecalling the raw data based on the type of library and flowcell.
--trim_strategy : Trim the adapter sequence present within each read. Activated by default. Passing "none" allow to prevent this behaviour.
--qscore_filtering : Allow to output reads in either "pass" or "fails" directories based on the mean quality score of the read. Deactivated by default.
--calib_detect : For detection and filtering of calibration strands. Deactivated by default.
-x : For activating GPU-based basecalling. "auto" allow to automatically detect the GPU of the machine onto which basecalling is performed.

After completion of the basecalling, all of the resulting fastq reads were gathered into a single fastq file for downstream processes.

b) Mapping reads onto *C. elegans* reference genome or transcriptome

All of the *C. elegans* reference files used for mapping of the sequencing reads were obtained from the wormbase release 270 (WS270). Complementary information regarding this version are available at: https://wormbase.org/about/wormbase_release_WS270#0--10

For mapping of nanopore reads onto C. elegans reference genome or transcriptome we used the minimap2 aligner, specifically designed for mapping long reads. Depending of the type of alignment, different settings were used, as indicated by the software documentation.
(available at: https://github.com/lh3/minimap2).

Genomic alignment:

```
minimap2 -ax splice -secondary=no -Y reference_genome.fa \
sequencing_reads.fastq > genome_alignment.sam
```

Transcriptomic alignment:

```
minimap2 -ax map-ont -secondary=no -Y reference_transcripts.fa \
sequencing_reads.fastq > transcriptome_alignment.sam
```

Description of the arguments:

-a : output alignment file in SAM (Sequence Alignment Map) format.
-x : preset for alignment mode. "splice" for splice-aware mapping (genome alignment) and "map-ont" for aligning reads in one block (transcriptome alignments).
-secondary : "no" prevents the generation of secondary alignments
-Y : prevents from trimming unaligned region on the extremities of the reads (hard-clipping).

c) Handling data and visualizing alignments

We used the Samtools suite for handling of the alignment files. This tool was designed for working with high-throughput sequencing data. It allows to import/export from the SAM format to perform different operations such as sorting, merging or efficiently retrieving reads mapping to a specific region.
(available at: http://www.htslib.org/doc/samtools.html)

Handling of alignment files through python is performed via the pysam library as further described in **section 4.4.2**.

Sorting and compression to BAM format:

The alignment file in SAM format was sorted for allowing downstream analysis. During this step, compression into a BAM (Binary Alignment Map) format was also performed by adding the ".bam" extension to the output file name.

```
samtools sort -o alignment.bam alignment.sam
```

Indexing:

Following sorting, the BAM file needs to be indexed. The index file allows for accessing reads mapped to a specific region or viewing alignments (see next section). The resulting file is generated automatically from the BAM file. The index files need to have the same name as the BAM file (with the ".bai" extension added to it) and be kept into the same directory.

```
samtools index alignment.bam
```

Visualization of the alignments:

After sorting and indexing, the alignments can be visualized with the Integrated Genome Viewer (IGV) software developed and maintained by the Broad Institute. This allow for performing visual inspection of the alignments and the general quality of the reads.
(Available at: http://software.broadinstitute.org/software/igv)

The reference file used during the alignment step needs to be provided to IGV for visualizing the different alignments. The alignment file in SAM/BAM format must be correctly sorted and indexed before being opened in IGV.

4.4.2 - Data analysis using python

During this project, and because of a lack of reference tools due to the recent development of nanopore sequencing, I wrote several scripts for performing the extraction and the analyse of the sequencing data obtained. All of the scripts were written in python language. The different libraries used and a description of the major scripts/algorithms developed are referenced in the following sub-sections.

The different python libraries used for developing various scripts are consigned in the **Table 8**.

Name	Usage	Available at
Python standard library		
random	generation of pseudo-random numbers	Included with python
re	regex operations	
functools	higher-order functions	
os	operating system interfaces	
Biological tools		
biopython	Set of tools for biological computation	https://biopython.org/
pysam	Manipulation of alignment files (SAM/BAM format)	https://github.com/pysam-developers/pysam
parasail-python	Python implementation of various sequence alignment algorithms	https://github.com/jeffdaily/parasail-python
Computing tools		
pandas	Manipulation of tabular data	https://pandas.pydata.org/
numpy	Standard library for working with numerical data	https://numpy.org/
scipy	Key algorithms and functions for scientific computing	https://www.scipy.org/
Visualization tools		
matplotlib	Standard plotting library to generate figures	https://matplotlib.org/
seaborn	Library for statistical data visualization	https://seaborn.pydata.org/
upsetplot	Library for visualiazing set overlaps	https://github.com/jnothman/UpSetPlot
DnaFeaturesViewer	Library for visualizing DNA features	https://github.com/Edinburgh-Genome-Foundry/DnaFeaturesViewer

Table 8 - Python libraries used during the project.

Annex 1

Bibliography

Ambros V. The functions of animal microRNAs. Nature. 2004;431(7006):350–5.

Anyanful A, Ono K, Johnsen RC, Ly H, Jensen V, Baillie DL, et al. The RNA-binding protein SUP-12 controls muscle-specific splicing of the ADF/cofilin pre-mRNA in C. elegans. J Cell Biology. 2004;167(4):639–47.

Bachvarova RF. A maternal tail of poly(a): The long and the short of it. Cell. 1992;69(6):895–7.

Banerjee AK. 5'-terminal cap structure in eucaryotic messenger ribonucleic acids. Microbiol Rev. 1980;44(2):175–205.

Bao Z, Murray JI, Boyle T, Ooi SL, Sandel MJ, Waterston RH. Automated cell lineage tracing in Caenorhabditis elegans. P Natl Acad Sci Usa. 2006;103(8):2707–12.

Barabino SM, Hübner W, Jenny A, Minvielle-Sebastia L, Keller W. The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. Gene Dev. 1997;11(13):1703–16.

Bell LR, Maine EM, Schedl P, Cline TW. Sex-lethal, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. Cell. 1988;55(6):1037–46.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–9.

Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc National Acad Sci. 1977;74(8):3171–5.

Berget SM, Sharp PA. A spliced sequence at the 5'-terminus of adenovirus late mRNA. Brookhaven Sym Biol. 1977;(29):332–44.

Beyer HM, Gonschorek P, Samodelov SL, Meier M, Weber W, Zurbriggen MD. AQUA Cloning: A Versatile and Simple Enzyme-Free Cloning Approach. Plos One. 2015;10(9):e0137652.

Bienroth S, Keller W, Wahle E. Assembly of a processive messenger RNA polyadenylation complex. Embo J. 1993;12(2):585–94.

Blencowe BJ. Alternative Splicing: New Insights from Global Analyses. Cell. 2006;126(1):37–47.

Blencowe BJ, Bowman JA, McCracken S, Rosonina E. SR-related proteins and the processing of messenger RNA precursors. Biochem Cell Biol. 1999;77(4):277–91.

Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, et al. A global analysis of Caenorhabditis elegans operons. Nature. 2002;417(6891):851–4.

Brenner S. The genetics of Caenorhabditis elegans. Genetics. 1974;77(1):71–94.

Bringmann P, Rinke J, Appel B, Reuter R, Lührmann R. Purification of snRNPs U1, U2, U4, U5 and U6 with 2,2,7-trimethylguanosine-specific antibody and definition of their constituent proteins reacting with anti-Sm and anti-(U1)RNP antisera. Embo J. 1983;2(7):1129–35.

Buratowski S, Hahn S, Guarente L, Sharp PA. Five intermediate complexes in transcription initiation by RNA polymerase II. Cell. 1989;56(4):549–61.

Burge CB, Tuschl T, Sharp PA. Splicing of Precursors to mRNAs by the Spliceosomes. In: Press CSHL, editor. The RNA World, Second Edition [Internet]. 1999 [cited 2020 Sep 12]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.692.2571&rep=rep1&type=pdf

BURNETT G, KENNEDY EP. The enzymatic phosphorylation of proteins. J Biological Chem. 1954;211(2):969–80.

Calarco JA, Xing Y, Cáceres M, Calarco JP, Xiao X, Pan Q, et al. Global analysis of alternative splicing differences between humans and chimpanzees. Gene Dev. 2007;21(22):2963–75.

Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357(6352):661–7.

Cassada RC, Russell RL. The dauerlarva, a post-embryonic developmental variant of the nematode Caenorhabditis elegans. Dev Biol. 1975;46(2):326–42.

Chalfie M, Tu Y, Euskirchen G, Ward W, Prasher D. Green fluorescent protein as a marker for gene expression. Science. 1994;263(5148):802–5.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell. 1977;12(1):1–8.

Colgan DF, Manley JL. Mechanism and regulation of mRNA polyadenylation. Gene Dev. 1997;11(21):2755–66.

Conrad R, Thomas J, Spieth J, Blumenthal T. Insertion of part of an intron into the 5' untranslated region of a Caenorhabditis elegans gene converts it into a trans-spliced gene. Mol Cell Biol. 1991;11(4):1921–6.

Consortium IHGS. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

Consortium* TC elegans S. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science. 1998;282(5396):2012–8.

Daneholt B. A Look at Messenger RNP Moving through the Nuclear Pore. Cell. 1997;88(5):585–8.

Deaton AM, Bird A. CpG islands and the regulation of transcription. Gene Dev. 2011;25(10):1010–22.

Dieci G, Fiorino G, Castelnuovo M, Teichmann M, Pagano A. The expanding RNA polymerase III transcriptome. Trends Genet. 2007;23(12):614–22.

Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284(5757):601–3.

Drygin D, Rice WG, Grummt I. The RNA Polymerase I Transcription Machinery: An Emerging Target for the Treatment of Cancer. Annu Rev Pharmacol. 2010;50(1):131–56.

Dupuy D, Bertin N, Hidalgo CA, Venkatesan K, Tu D, Lee D, et al. Genome-scale analysis of in vivo spatiotemporal promoter activity in Caenorhabditis elegans. Nat Biotechnol. 2007;25(6):663–8.

Dupuy D, Li Q-R, Deplancke B, Boxem M, Hao T, Lamesch P, et al. A First Version of the Caenorhabditis elegans Promoterome. Genome Res. 2004;14(10b):2169–75.

Dvir A, Conaway JW, Conaway RC. Mechanism of transcription initiation and promoter escape by RNA polymerase II. Curr Opin Genet Dev. 2001;11(2):209–14.

Ebbing A, Vértesy Á, Betist MC, Spanjaard B, Junker JP, Berezikov E, et al. Spatial Transcriptomics of C. elegans Males and Hermaphrodites Identifies Sex-Specific Differences in Gene Expression Patterns. Dev Cell. 2018;47(6):801-813.e6.

Edwalds-Gilbert G, Veraldi KL, Milcarek C. Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res. 1997;25(13):2547–61.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009;323(5910):133–8.

Ferguson KC, Heid PJ, Rothman JH. The SL1 trans-spliced leader RNA performs an essential embryonic function in Caenorhabditis elegans that can also be supplied by SL2 RNA. Gene Dev. 1996;10(12):1543–56.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998;391(6669):806–11.

Gallie DR. The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. Gene Dev. 1991;5(11):2108–16.

Gerstein MB, Lu ZJ, Nostrand ELV, Cheng C, Arshinoff BI, Liu T, et al. Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. Science. 2010;330(6012):1775–87.

Giordano-Santini R, Dupuy D. Selectable genetic markers for nematode transgenesis. Cell Mol Life Sci. 2011;68(11):1917–27.

Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD. Structural Basis of Transcription: An RNA Polymerase II Elongation Complex at 3.3 Å Resolution. Science. 2001;292(5523):1876–82.

Gómez-Saldivar G, Osuna-Luque J, Semple JI, Glauser DA, Jarriault S, Meister P. Tissue-specific transcription footprinting using RNA PoI DamID (RAPID) in C. elegans. Biorxiv. 2020;2020.08.19.257873.

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. Science. 2005;308(5720):419–21.

Hammarlund M, Hobert O, Miller DM, Sestan N. The CeNGEN Project: The Complete Gene Expression Map of an Entire Nervous System. Neuron. 2018;99(3):430–3.

Hashimoto C, Steitz JA. A small nuclear ribonucleoprotein associates with the AAUAAA polyadenylation signal in vitro. Cell. 1986;45(4):581–91.

Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. RNA Polymerase IV Directs Silencing of Endogenous DNA. Science. 2005;308(5718):118–20.

Hir HL, Gatfield D, Izaurralde E, Moore MJ. The exon–exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. Embo J. 2001;20(17):4987–97.

Holde KE van. Chromatin. J Mol Recognit. 1988;2(3):i–i.

Hunter T. Protein kinases and phosphatases: The Yin and Yang of protein phosphorylation and signaling. Cell. 1995;80(2):225–36.

Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, Fang L, et al. High-Throughput In Vivo Analysis of Gene Expression in Caenorhabditis elegans. Plos Biol. 2007;5(9):e237.

Ibach J, Brakmann S. Sequencing Single DNA Molecules in Real Time. Angewandte Chemie Int Ed. 2009;48(26):4683–5.

Ishigaki Y, Li X, Serin G, Maquat LE. Evidence for a Pioneer Round of mRNA Translation mRNAs Subject to Nonsense-Mediated Decay in Mammalian Cells Are Bound by CBP80 and CBP20. Cell. 2001;106(5):607–17.

Jeffreys AJ, Wilson V, Thein SL. Individual-specific 'fingerprints' of human DNA. Nature. 1985;316(6023):76–9.

Jenuwein T, Allis CD. Translating the Histone Code. Science. 2001;293(5532):1074–80.

Jeong S. SR Proteins: Binders, Regulators, and Connectors of RNA. Mol Cells. 2017;40(1):1–9.

Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 2015;43(21):e143–e143.

Keiper BD, Lamphear BJ, Deshpande AM, Jankowska-Anyszka M, Aamodt EJ, Blumenthal T, et al. Functional Characterization of Five eIF4E Isoforms inCaenorhabditis elegans. J Biol Chem. 2000;275(14):10590–6.

Kent WJ, Zahler AM. The Intronerator: exploring introns and alternative splicing in Caenorhabditis elegans. Nucleic Acids Res. 2000;28(1):91–3.

Krause M, Hirsh D. A trans-spliced leader sequence on actin mRNA in C. elegans. Cell. 1987;49(6):753–61.

Kuroyanagi H, Kobayashi T, Mitani S, Hagiwara M. Transgenic alternative-splicing reporters reveal tissue-specific expression profiles and regulation mechanisms in vivo. Nat Methods. 2006;3(11):909–15.

Kuroyanagi H, Ohno G, Sakane H, Maruoka H, Hagiwara M. Visualization and genetic analysis of alternative splicing regulation in vivo using fluorescence reporters in transgenic Caenorhabditis elegans. Nat Protoc. 2010;5(9):1495–517.

Lall S, Friedman CC, Jankowska-Anyszka M, Stepinski J, Darzynkiewicz E, Davis RE. 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. J Biol Chem. 2004;279(44):45573–85.

Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843–54.

Lewis JD, Gunderson SI, Mattaj IW. The influence of 5′ and 3′ end structures on pre-mRNA metabolism. J Cell Sci. 1995;1995(Supplement 19):13–9.

Liou RF, Blumenthal T. trans-spliced Caenorhabditis elegans mRNAs retain trimethylguanosine caps. Mol Cell Biol. 1990;10(4):1764–8.

Luo J, Ying K, Bai J. Savitzky–Golay smoothing and differentiation filter for even number data. Signal Process. 2005;85(7):1429–34.

Luo M, Reed R. Splicing is required for rapid and efficient mRNA export in metazoans. Proc National Acad Sci. 1999;96(26):14937–42.

Maddams WF, Mead WL. The measurement of derivative i.r. spectra—I. Background studies. Spectrochimica Acta Part Mol Spectrosc. 1982;38(4):437–44.

Maroney PA, Denker JA, Darzynkiewicz E, Laneve R, Nilsen TW. Most mRNAs in the nematode Ascaris lumbricoides are trans-spliced: a role for spliced leader addition in translational efficiency. Rna New York N Y. 1995;1(7):714–23.

Maxam AM, Gilbert W. A new method for sequencing DNA. Proc National Acad Sci. 1977;74(2):560–4.

Mayeda A, Krainer AR. Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. Cell. 1992;68(2):365–75.

Mayeda A, Munroe SH, Cáceres JF, Krainer AR. Function of conserved domains of hnRNP A1 and other hnRNP A/B proteins. Embo J. 1994;13(22):5483–95.

Mello CC, Kramer JM, Stinchcomb D, Ambros V. Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. Embo J. 1991;10(12):3959–70.

Minvielle-Sebastia L, Keller W. mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. Curr Opin Cell Biol. 1999;11(3):352–7.

Moore MJ, Sharp PA. Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. Nature. 1993;365(6444):364–8.

Nussinov R, Tsai C-J, Xin F, Radivojac P. Allosteric post-translational modification codes. Trends Biochem Sci. 2012;37(10):447–55.

Orgel LE, Crick FHC. Selfish DNA: the ultimate parasite. Nature. 1980;284(5757):604–7.

Padgett R, Konarska M, Grabowski P, Hardy S, Sharp P. Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. Science. 1984;225(4665):898–903.

Praitis V, Casey E, Collar D, Austin J. Creation of low-copy integrated transgenic lines in Caenorhabditis elegans. Genetics. 2001;157(3):1217–26.

Preußer C, Jaé N, Bindereif A. mRNA splicing in trypanosomes. Int J Med Microbiol. 2012;302(4–5):221–4.

Ramani AK, Calarco JA, Pan Q, Mavandadi S, Wang Y, Nelson AC, et al. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. Genome Res. 2011;21(2):342–8.

Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. The full-length transcriptome of C. elegans using direct RNA sequencing. Genome Res. 2020;30(2):299–312.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc National Acad Sci. 1977;74(12):5463–7.

Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochem J. 1951;49(4):463–81.

Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem. 1964;36(8):1627–39.

Seydoux G, Fire A. Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. Dev Camb Engl. 1994;120(10):2823–34.

Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–95.

Shen Z, Zhang X, Chai Y, Zhu Z, Yi P, Feng G, et al. Conditional Knockouts Generated by Engineered CRISPR-Cas9 Endonuclease Reveal the Roles of Coronin in C. elegans Neural Development. Dev Cell. 2014;30(5):625–36.

Singh J, Padgett RA. Rates of in situ transcription and splicing in large human genes. Nat Struct Mol Biol. 2009;16(11):1128–33.

Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T. Operons in C. elegans: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell. 1993;73(3):521–32.

Steinmann-Zwicky M, Amrein H, Nöthiger R. Genetic Control Of Sex Determination In Drosophila. Adv Genet. 1990;27:189–237.

Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol. 1977;56(1):110–56.

Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol. 1983;100(1):64–119.

Sutton RE, Boothroyd JC. Evidence for Trans splicing in trypanosomes. Cell. 1986;47(4):527–35.

TAMURA T-A, KONISHI Y, MAKINO Y, MIKOSHIBA K. MECHANISMS OF TRANSCRIPTIONAL REGULATION AND NEURAL GENE EXPRESSION**Part of this communication was presented at the Group Dinner Conference on Gene Transcription, which was organized by Y. Yoneda as part of the 38th Meeting of the Japanese Society for Neurochemistry held on 1 July 1995 in Kyoto, Japan. Y. Yoneda also acted as executive editor in the refereeing process of this manuscript. Neurochem Int. 1996;29(6):573–81.

Tan JH, Fraser AG. The combinatorial control of alternative splicing in C. elegans. Plos Genet. 2017;13(11):e1007033.

Tanii T, Akahori R, Higano S, Okubo K, Yamamoto H, Ueno T, et al. Improving zero-mode waveguide structure for enhancing signal-to-noise ratio of real-time single-molecule fluorescence imaging: A computational study. Phys Rev E. 2013;88(1):012727.

Thomas JD, Conrad RC, Blumenthal T. The C. elegans Trans-spliced leader RNA is bound to Sm and has a trimethylguanosine cap. Cell. 1988;54(4):533–9.

Tintori SC, Osborne Nishimura E, Golden P, Lieb JD, Goldstein B. A Transcriptional Lineage of the Early C. elegans Embryo. Dev Cell. 2016;38(4):430–44.

Tourasse NJ, Millet JRM, Dupuy D. Quantitative RNA-seq meta-analysis of alternative exon usage in C. elegans. Genome Res. 2017;27(12):2120–8.

Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science. 2010;329(5992):689–93.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72.

Wahle E, Lustig A, Jenö P, Maurer P. Mammalian poly(A)-binding protein II. Physical properties and binding to polynucleotides. J Biological Chem. 1993;268(4):2937–45.

Walter MC, Zwirglmaier K, Vette P, Holowachuk SA, Stoecker K, Genzel GH, et al. MinION as part of a biomedical rapidly deployable laboratory. J Biotechnol. 2017;250:16–22.

Wang Z, Burge CB. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. Rna. 2008;14(5):802–13.

Watabe E, Ono S, Kuroyanagi H. Alternative splicing of the Caenorhabditis elegans lev-11 tropomyosin gene is regulated in a tissue-specific manner. Cytoskeleton. 2018;75(10):427–36.

WATSON JD, CRICK FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171(4356):737–8.

Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc National Acad Sci. 2011;108(Supplement 2):10863–70.

Wong S. Diagnostics in space: will zero gravity add weight to new advances? Expert Rev Mol Diagn. 2019;20(1):1–4.

Yonaha M, Proudfoot NJ. Transcriptional termination and coupled polyadenylation in vitro. Embo J. 2000;19(14):3770–7.

Zhou Z, Luo M, Straesser K, Katahira J, Hurt E, Reed R. The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature. 2000;407(6802):401–5.

Zhu P, Craighead HG. Zero-Mode Waveguides for Single-Molecule Analysis. Annu Rev Biophys. 2012;41(1):269–93.

Zorio DAR, Cheng NN, Blumenthal T, Spieth J. Operons as a common form of chromosomal organization in C. elegans. Nature. 1994;372(6503):270–2.

Investigation of *Caenorhabditis elegans* transcriptome using nanopore-based sequencing technology.

A recent meta-analysis of alternative exon usage in *C. elegans* refined our comprehension of its transcriptome, especially regarding the splicing quantitative aspects of alternative splicing in messenger RNAs. However, Next-Generation Sequencing technologies like Illumina technology are proving to be limited to fully characterize one's transcriptome. PCR-based sequencing methods are known to introduce amplification bias affecting the overall distribution of mRNAs detected in one experiment and short-reads are not suited to accurately predict the frequency of isoforms derived from multiple alternative splicing events.

In this study, we exploited new possibilities offered by Oxford Nanopore Technology (ONT) to overcome those limitations. Nanopore-based sequencing allow us to directly sequence nucleic acids without any prior amplification step and generates long-reads covering up to the full-length of the molecule. Hence, permitting to further characterize *C. elegans* transcriptome by providing a more accurate measure of isoforms ratios, a better comprehension of exons associations during alternative splicing and by characterizing differentially trans-spliced mRNAs.

We assessed the efficiency of different sequencing kits commercialized by ONT and our results indicates that direct-cDNA sequencing is better suited for performing transcriptome analysis in *C. elegans*, in regard to the quantity and quality of data generated. Following this analysis, several direct-cDNA sequencing experiments have been performed on different populations of mRNAs: libraries of poly(A) RNAs representing the whole-animal transcriptome and libraries of SL1-enriched mRNAs. Our findings indicates that trans-spliced RNAs have an atypical behaviour during ONT's library preparation and trans-splicing of *C. elegans* mRNAs is more prevalent than previously reported. Finally, we also show that alternative promoters can lead to population of isoformes exhibiting different trans-splicing status.

Keywords: Caenorhabditis elegans ; Oxford Nanopore Technology ; RNA sequencing ; alternative splicing ; trans-splicing.

Etude du transcriptome de *Caenorhabditis elegans* par la technique de séquencage nanopore.

Une récente méta-analyse de l'utilisation des jonctions exon-exon chez C. elegans nous a permis d'améliorer notre compréhension de son transcriptome, en particulier pour la quantification des événenements d'épissage alternatif au sein des ARNs messagers. Cependant, les technologies de séquençage NGS, comme l'Illumina, se montrent limitées pour l'étude des transcriptomes. Ces technologies, basées sur la PCR, entraînent l'apparition de biais d'amplifications qui affectent la distribution des ARNms détectés au sein d'une expérience, et la petite taille des fragments générés n'est pas adapté pour correctement déterminer la fréquence des isoformes issus des différents événements d'épissage alternatifs.

Dans cette étude, nous avons utilisé les nouvelles possibilités offertes par les séquenceurs Oxford Nanopore Technology (ONT) pour nous affranchir de ces limitations. Le séquençage nanopore nous permet de séquencer directement des molécules d'acides nucléiques sans avoir recours à des étapes d'amplification, et permet de générer des fragments pouvant couvrir la taille totale de la molécule séquencée. Ceci nous permettant de mieux caractériser le transcriptome de *C. elegans*, en effectuant une mesure plus précise des fréquences d'isoformes, et en nous permettant une meilleure compréhension des événements d'épissage alternatif et de trans-épissage qui affectent les messagers.

Nous avons évalué trois différents kits de séquençage commercialisés par ONT et nos résultats suggèrent que le kit de séquençage direct-cDNA est plus adapté aux études de transcriptomes, en termes de quantité et de qualité de données générées. À la suite de cette analyse, plusieurs séquençages direct-cDNA ont été réalisés sur différentes populations d'ARNs : des libraries d'ARNs poly(A) représentants le transcriptome entier de *C. elegans*, et des libraries enrichies en ARNs SL1. Nos résultats indiquent que les ARNs trans-épissés se comportent différemment lors de la préparation des libraries ONT et que les phénomènes de trans-épissage sont plus prédominants que précédemment rapporté. Enfin, nous montrons que les promoteurs alternatifs peuvent conduire à la génération de populations d'isoformes présentant différents statuts de trans-épissage.

Mots-clés : Caenorhabditis elegans ; Oxford Nanopore Technology ; séquencage ARNs ; épissage alternatif ; trans-épissage.