Functional RNA-RNA interactions in the context of circular RNAs

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)
by

Stefan Stefanov

submitted to the Department of Biology, Chemistry,
Pharmacy
of Freie Universität Berlin
28.02.2023

Institute: Max Delbrück Center-Berlin

Time Period: November 2016 – February 2023

Supervisor: Prof. Dr. Irmtraud Meyer

1st Reviewer: Prof. Dr. Irmtraud Meyer

2nd Reviewer: Prof. Dr. Knut Reinert

3rd Reviewer: Prof. Dr. Dominik Seelow

Date of defence: 17.04.2024

Declaration of Independence

Herewith I certify that I have prepared and written my thesis independently and that I have not used any sources and aids other than those indicated by me.
This dissertation has not yet been presented to any other examination authority in the same or a similar form and has not yet been published.

Outline

Definitions

AS Alternative splicing – the alternative combination of exons during the processing of the nascent RNA molecule

BSJ Back-splice junction – the point of aberrant splicing leading to a circular RNA molecule

circRNA circular RNA

circRNA enriched library – RNA-seq library that has undergone an enrichment procedure for cirRNAs

DEA Differential expression analysis

dNTP deoxyribonucleoside triphosphate

dsRNA Double stranded RNA – an RNA duplex formed within a single RNA molecule or 2 separate RNA molecules

EM Expectation-Maximisation – in the case of transcriptomics an iterative procedure to optimise abundance estimation

FDR False discovery rate

FSJ Forward-splice junction – a canonical splice junction

GO Gene ontology

HERVHhigh cells – cell line selected from hESC based on high levels HERVH promoter activity

KD – knockdown of a gene product; targeted decrease of transcript levels with the intention of identifying the function of the gene/transcript

MSigDB Molecular Signatures Database

NSG Next-generation sequencing

RRI RNA-RNA interaction – trans forming RNA duplex between two separate RNA molecules

Chapter 1

Summary

1.1 English

The purpose of my project is to identify novel functions of circRNAs with a particular focus on the effects of RNA–RNA interactions (RRI) on RNA processing. Computational prediction of RRI has revealed the biological function and mechanism of action of multiple genes. However, computational RRI prediction is limited by 2 major challenges: knowing the full sequence of the transcript and a high false positive rate. Discovering the full sequence identity of circRNA has been a challenging task for bioinformaticians in the last decade. In addition, the lack of knowledge of the full sequence of the transcripts in a sample leads to skewed quantification based on RNA-seq data, as well as incorrect results from analyses of NGS-derived techniques (e.g. CLIP-seq, SPLASH etc.). The problem of false discovery of new RRIs can be mitigated by dedicated experimental datasets.

To overcome the first hurdle of my project, I developed *CYCLeR*, a computational tool that compares ribo-depleted and circRNA enriched RNA-seq libraries and outputs a high-confidence set of circRNA transcripts. The true strength of *CYCLeR* is the quantification module that can robustly estimate the abundances of both circular and linear transcripts. I have shown the advantage of *CYCLeR* over alternative tools in terms of transcript assembly and quantification. I have also shown that *CYCLeR* has is the only tool suitable to search for the functional association of circRNA transcripts.

The second second part of my work focuses on predicting functional RRIs that influence pluripotency. A co-expression network based on the output of *CYCLeR* can show the association of circRNA with known biological pathways and significantly facilitate the discovery of the function of circRNA. In vivo RNA proximity ligation experiments provide information on the dynamics of RNA-RNA interaction inside the cell. The combination of RNA-seq and RNA interactome data allows me to significantly enhance the strength of computational predictions.

I build a co-expression network based on time series experiment of H1ESC treated with retinoic acid. I combine the co-expression information with results from analysis of RNA-RNA proximity ligation data (SPLASH). The analysis is supplemented with localisation information based on RNA-seq libraries specific for nuclear localisation. The results two circRNAs that participate in functional RRIs.

circFIRRE is significantly enriched in SPLASH data, indicating a high probability of interaction with other RNAs. Interestingly, circFIRRE is one of the few circRNAs specifically enriched in the nucleus. The enrichment can be explained by the binding site for the hnRNPU protein, which keeps the circRNA in the nucleus. Knockout of the circFIRRE locus in human leads to a viral response. Multiple interaction sites of circFIRRE with ALU-specific sequences indicate that the viral response is triggered by disruption of A-to-I editing in cells.

circLARP7 is another nuclear-specific circRNA. circLARP7 is co-expressed with all major markers for pluripotency. It is also expressed in high proximity to MIR302CHG – a microRNA host gene related to maintaining the pluripotent state. High complementarity and conservation of a duplex between the circLARP7 and the nascent MIR302CHG indicate that circLARP7 might be related to the processing of the microRNAs from the miR-302/367 cluster.

1.2 Deutsch

Das Ziel meines Projekts ist es, neue Funktionen von circRNAs zu identifizieren, mit besonderem Fokus auf die Auswirkungen von RNA–RNA-Interaktionen (RRI) auf die RNA-Verarbeitung. Die computergestützte Vorhersage von RRI hat die biologische Funktion und den Wirkungsmechanismus mehrerer Gene offenbart. Jedoch wird die Vorhersage von RRI durch zwei wesentliche Herausforderungen beschränkt: die Kenntnis der vollständigen Sequenz des Transkripts und eine hohe falsch-positive Rate. Die Aufschlüsselung der vollständen Sequenz von circRNA stellte in den letzten zehn Jahren eine große Herausforderung für Bioinformatiker dar. Darüber hinaus führt die mangelnde Kenntnis der vollständigen Sequenz der Transkripte in einer Probe zu einer verzerrten Quantifizierung auf der Grundlage von RNA-seq-Daten sowie zu falschen Ergebnissen aus Analysen von NGS-abgeleiteten Techniken (z. B. CLIP-seq, SPLASH usw.). Das Problem einer hohen Falscherkennungsrate neuer RRIs kann durch Nutzung geeigneter experimenteller Datensätze begrenzt werden.

Um die erste Hürde meines Projekts zu überwinden, habe ich CYCLeR entwickelt, ein Computertool, das Ribo-abgereicherte und circRNA-angereicherte RNA-seq-Bibliotheken vergleicht und einen Reihe von circRNA-Transkripten mit hoher Zuverlässigkeit ausgibt. Die wahre Stärke von CYCLeR ist das Quantifizierungsmodul, das die Häufigkeit von sowohl kreisförmigen als auch linearen Transkripten zuverlässig berechnen kann. Ich habe den Vorteil von CYCLeR gegenüber alternativen Tools in Bezug auf Transkript-Zusammenstellung und Quantifizierung aufgezeigt. Ich habe auch gezeigt, dass CYCLeR das einzige geeignete Werkzeug ist, um nach der funktionellen Verbindung von circRNA-Transkripten zu suchen.

Der zweite Teil meiner Arbeit konzentriert sich auf die Vorhersage funktioneller RRIs, die die Pluripotenz beeinflussen. Ein auf der Ausgabe von CYCLeR basierendes Koexpressionsnetzwerk kann die Verbindung von circRNA mit bekannten biologischen Signalwegen aufzeigen und die Entdeckung der Funktion von circRNA erheblich erleichtern. In-vivo-RNA-Proximity-Ligation-Experimente liefern Informationen über die Dynamik der RNA-RNA-Interaktion innerhalb der Zelle. Die Kombination von RNA-Seq- und

RNA-Interaktom-Daten ermöglicht es mir, die Aussagekraft von Computervorhersagen erheblich zu verbessern.

Ich baue ein Koexpressionsnetzwerk basierend auf einem longitudinalen Experiment mit H1ESC Zellen, welche mit Retinsäure behandelt wurden und kombiniere die Koexpressionsinformationen mit Ergebnissen aus der Analyse von RNA-RNA-Proximity-Ligation-Daten (SPLASH). Die Analyse wird durch Lokalisierungsinformationen basierend auf RNA-seq-Bibliotheken ergänzt, die für die Kernlokalisierung spezifisch sind. Die Ergebnisse weisen auf zwei circRNAs hin, die an funktionellen RRIs beteiligt sind.

circFIRRE ist in SPLASH-Daten signifikant angereichert, was auf eine hohe Wahrscheinlichkeit einer Wechselwirkung mit anderen RNAs hinweist. Interessanterweise ist circFIRRE eine der wenigen circRNAs, die spezifisch im Zellkern angereichert sind, was sich mit der Bindungsstelle für das hnRNPU-Protein erklären lässt, das die circRNA im Zellkern hält. Der Knockout des circFIRRE-Locus im Menschen führt zu einer viralen Reaktion. Mehrere Interaktionsstellen von circFIRRE mit ALU-spezifischen Sequenzen weisen darauf hin, dass die virale Reaktion durch Unterbrechung der A-zu-I-Editierung in Zellen ausgelöst wird.

circLARP7 ist eine weitere kernspezifische circRNA und wird mit allen wichtigen Markern für Pluripotenz koexprimiert. Es wird auch in großer Nähe zu MIR302CHG exprimiert – einem Mikro-RNA-Wirtsgen, das mit der Aufrechterhaltung des pluripotenten Zustands in Zusammenhang steht. Hohe Komplementarität und Konservierung eines Duplex zwischen dem circLARP7 und dem entstehenden MIR302CHG deuten darauf hin, dass circLARP7 mit der Prozessierung der microRNAs aus dem miR-302/367-Cluster zusammenhängen könnte.

Chapter 2

Introduction

The purpose of this chapter is to introduce the reader to the nature and aims of the project and provide the background knowledge necessary for the comprehension of the work described in the following chapters.

2.1 Aims of the project

The goal of my project is to identify circRNAs that participate in functional RRIs that manage pluripotency. The functional search for RRIs requires knowledge of the sequence of the transcript and experimental data that would narrow the search space. My work can be separated into two distinct aims:

2.1.1 Develop a novel tool for circular transcript assembly and quantification

The objective of a transcriptomic analysis is to identify the full sequence of every (linear or circular) transcript in a sample and its relative abundance. CircRNAs pose a particular challenge due to their low abundance and incompatibility with methods designed for the assembly of linear RNA. I aim to fill this gap in the field by presenting my conceptually novel tool *CYCLeR* (Co-estimate Your Circular and Linear RNAs).

2.1.2 Study the circRNAs effect on pluripotency based on a dedicated dataset

Computational prediction of functional RNA structures and interactions has revealed the functions of many non-coding genes. But the challenges of computational RRI prediction require limiting the number of potential candidates. I can determine the likely circRNA candidates affecting pluripotency by building a co-expression network based on time series data of human ESC differentiation. By combining the co-expression information with transcript localisation, interactome, and editome data, I can narrow down the set of potential functional circRNAs significantly. With the subsequent computational RRI prediction, I can add nucleotide level precision of the interaction. I aim to use the combination of experimental and computational methods to precisely predict the biological function of a circRNA-RNA interaction.

2.2 Splicing

Expression of the eukariotic gene happens when an RNA molecule is transcribed from the DNA template. Splicing is the process of transforming the newly synthesised RNA into a mature RNA. Splicing satisfies the need for alternative products with small differences between them, depending on the need of the cell [1]. The unspliced molecule is referred to as nascent transcript and pre-RNA, and the finished product is termed mature RNA. Splicing occurs by removing the so-called introns from the nascent RNA molecule [2]. The selection of introns to be removed provides diversity in the splicing of an RNA molecule. Different combinations of excluded intron sequences lead to alternative splicing and alternative products of the transcribed gene(s)–isoforms. The regions of the nascent RNA that remain are called exons. The set of exons in the final RNA molecule can differ due to splicing that removes a portion or even a full exon.

Since most RNAs are transcribed in the nucleus, splicing is considered a process that is primarily localised in the nucleus. There are different mechanisms of splicing. The most common form of splicing in eukaryotes is mediated by the spliceosome [3] – a complex of RNA (snRNAs) and proteins. The spliceosome complex is divided into two types: the major spliceosome and the minor [4]. The major one targets intronic sequences, including GU at the 5' splice site and AG at the 3' splice site. The minor spliceosome targets a variety of intron motifs that are less common. Some introns have the ability to self-splice; however, that is a rare event compared to spliceosome-mediated splicing [5]. tRNAs also undergo splicing, but their biochemical mechanism differs significantly [6].

Spliceosome-mediated splicing and self-splicing occur in a similar manner. Biochemically, the process is based on two-step transesterification reactions. In the first step, the 2' hydroxyl group (2'-OH) of the branch point nucleotide adenosine attacks the phosphate at the 5' exon–intron junction (5' splice site), resulting in the cleavage of the phosphodiester bond between the 5' exon and intron and the concurrent formation of a new 5'–2' phosphodiester bond between the 5'end of the intron and the branch point adenosine. Thus, a lariat-structured intermediate (lariat intron-3' exon) and a cut-off 5' exon intermediate are produced. In the second step, the 3'-OH group of the cut-off 5' exon attacks the phosphate at the intron - 3' exon junction (3' splice site), releasing the lariat intron product and generating the spliced mature mRNA product [7].

Splicing is mediated by a combination of elements within the molecule undergoing splicing (cis-acting element) and different molecules (trans-acting elements). The trans-acting element is usually a protein, but can also be a trans-acting RNA. The trans-acting element can bind to a specific sequence (cis-regulatory element) in the nascent RNA molecule that is either a splicing enhancer or a repressor. The regulation target site can be either exon or intron. Trans-acting elements can also bind directly to a splice site, making it inaccessible. RNA structure can also influence splicing by making the splice site or regulatory sites inaccessible or, the opposite, ensuring splicing by bringing the splice sites in proximity [3]. The speed of RNA transcription influences the splicing process by allowing a limited time for the splicing factors to bind to their targets.

Splicing can occur both co-transcriptionally and post-transcriptionally [8, 9]. The sole limitation is the accessibility of the splice sites and the splice control sites in the sequence. Splicing can take place in forward fashion - an upstream 5' site interacting with a downstream 3' site, or alternatively, a downstream 5' site interacting with an upstream 3' site (see Figure 2.1). The latter leads to a circular RNA molecule. Splicing does not occur exclusively within one gene. Sometimes, nascent RNA from one gene can splice with nascent RNA from another gene, forming a trans-splicing product.

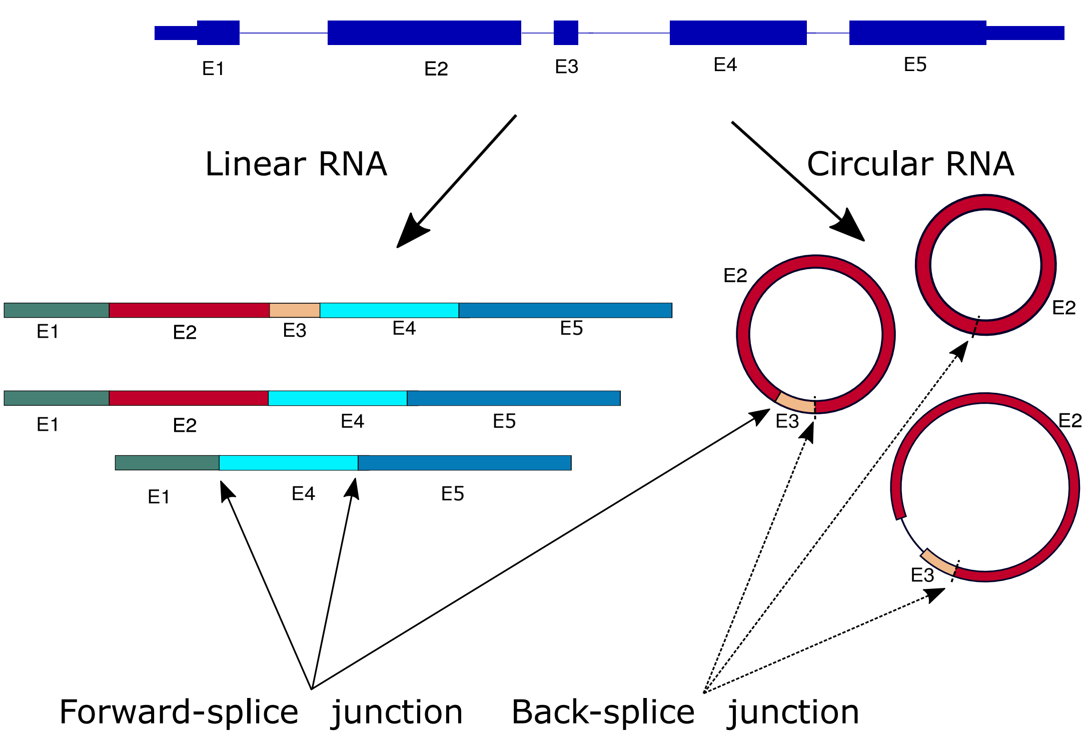


Figure 2.1: **Splicing** *Diversity of splicing isoforms*

2.2.1 Forward-splicing

To reiterate, forward-splicing occurs when an upstream 5' splice site interacts with a downstream 3' splice site. The generally accepted forward alternative splicing event terms can be summarised as follows [3]:

- Exon skipping – exon being completely spliced out of the final transcript
- Mutually exclusive exons – just one of a pair of exons can remain in the final molecule
- Alternative 5' splicing – an alternative donor site of an exon is used
- Alternative 3' splicing – an alternative acceptor site of an exon is used
- Intron retention – a sequence can be spliced out as an intron is retained instead

Alternative splicing (AS) events differ in rates between organisms. For example, in mammals, the most common type of AS is exon skipping, and the least common is intron retention. In contrast, in plants, intron retention is the most common event [10].

2.2.2 Back-splicing

While forward splicing is the predominant mode of splicing, in certain cases, a downstream 5‘ donor site connects with an exon with an upstream 3’ acceptor site, producing a circular RNA molecule (circRNA) [11, 12]. Such splicing is called head-to-tail splicing or back-splicing. The irregular junction site is called a back-splice junction (BSJ). For a long period of time, scientists have focused only on linear RNA transcripts, but in recent years research of circular RNAs (circRNAs) has emerged as a new field [11, 12, 13]. Circular RNAs were initially dismissed as naturally occurring errors and simply side effects of forward-splicing. Even after the presence of circular RNA was proven via electron microscopy [14], their importance was diminished due to the fact that no apparent function was identified. With the discovery of circRNA, flanking introns are shown to be imperative for circularisation [14]. The presence of ALU elements and protein binding has been shown to facilitate circularisation [9]. Circular molecules are also produced as part of the lariat intermediate during splicing. If the debranching enzyme does not process the lariat structure, it is assumed that exonucleases degrade the 3’ tail of the lariat and it remains in the cell as a circular molecule [15, 16]. Quantitatively, circular transcripts amount to 1-3% of linear poly-A transcripts. There is no evidence to suggest that circular splicing is a by-product of linear splicing [13].

2.2.3 Functional roles of circular splicing isoforms

Enrichment of circular RNAs in specific tissues indicates that there is, in fact, a controlled mechanism of their production. Circular RNAs, such as those from the *Sry* gene, accumulate in mouse testes with relatively high abundance. Circular RNAs are relatively more abundant in brain tissue, a trait is consistent across species. Some circular RNAs (e.g. circSry , circCDR1-AS) have been proposed to act as microRNA sponges [12, 13]. Others are suggested to control the expression of the genes from which they are derived- circMbl in Drosophila [9]; circEIF3J and circPAIP2 in human [17]. There is even proof that circular RNAs can be translated to produce a protein product driven by N6-methyladenosine [18]. Although the proposed method for circRNAs synthesis would suggest their presence predominantly in the nucleus, circular RNAs tend to be enriched in the cytoplasm [11, 13].

2.3 Transcriptome Sequencing

2.3.1 Background

The majority of research regarding the transcriptome of the cell views RNA as a simple intermediary for protein synthesis. Scientists often study the variety and levels of the transcript as an indirect indicator of the state of the proteome. High-throughput sequencing technology has provided great insight into the transcriptome activity of the cell and has triggered the discovery of novel isoforms and functions of RNA.

2.3.2 Next-generation sequencing

A turning point in transcriptome study was the development of Next-generation sequencing (NGS), also known as Massive parallel sequencing. After the development of alternative approaches for high-throughput sequencing, the term NGS is often substituted with Second-generation sequencing.

Although NGS strategies across different platforms involve different chemistry and require distinct technical configurations, the common objective of all NGS approaches is the identification of the sequence of a short DNA fragment. NGS is an alternative to Sanger (First-generation) sequencing that is based on of "chain-termination". In this method, the polymerisation of a DNA polymerisation chain reaction is blocked chemically with di-deoxynucleotide triphosphates, and the alternative products are separated by size and analysed [19, 20]. The capability to process a massive number of DNA fragments in parallel has allowed the expansion of NGS into RNA-specific studies.

The RNA sequencing (RNA-seq) uses NGS technology to provide qualitative and quantitative information on the transcriptome of the cell [21, 22]. RNA is used to produce an ensemble of cDNA fragments, making a "library". These fragments are sequenced, and through multiple computational procedures, the sequencing information can recreate the sequence and relative abundance of the

RNA transcripts. There is a limitation of the NGS technology fragment size. When sequencing fragments are longer than 700 nucleotides (nts), the error rate of base calling substantially increases. The implications of this limit force the need to study RNA not as a whole molecule, but instead as a set of cDNA fragments.

Before the widespread adoption of RNA-seq, the predominant technology used to study transcript levels of expression was microarray. In microarray technology, fluorescently-labeled targets are introduced to a chip with multiple spots where specific hybridisation can occur to probes. The level of probe-to-target hybridisation is measured as fluorescence intensity [23, 24]. The major difference between microarray and RNA-seq technology is the variance in resulting measurement. While microarray results are generally measured as a single target leading to a single signal, the RNA-seq results trace multiple fragments to a single molecule. Therefore, microarray data benefit from stable variance-to-mean ratio between measurements for different probes, while RNA-seq results vary based on the mean abundance of an RNA and lead to the so called overdispersion [25]. This means that the computational procedures that were designed for microarray data cannot be readily applied to RNA-seq data. The major advantage of RNA-seq results is the option to focus on unannotated isoforms, as opposed to limiting the study to a set of pre-selected probes. In fact, with the addition of novel isoform information, old RNA-seq data can be re-analysed, leading to an improvement in the results.

2.3.2.1 RNA enrichment

The RNA content in a prokaryotic or eukaryotic cell consists of 80–90% ribosomal RNA (rRNA), 10–15% tRNA and 3–7% messenger RNA and regulatory ncRNA [26]. Most studies focus primarily on mRNA and ncRNA. The high levels of rRNA and tRNA from the bulk RNA extract produce a correspondingly high level of library fragments, which sparks very little scientific interest. It is cost- and effort-efficient to enrich the RNA extract for RNAs of interest. Ironically, most of the so-called enrichment procedures are, in fact, depletion procedures against a specific type of RNA. The common enrichment procedures can be summarised as:

size selection tRNAs are usually filtered out during a standard size selection of 200 nt.

poly(A) enrichment It is very common to use trascriptomic data as a proxy for protein level analysis. In such cases, mRNA transcripts a selected with oligo(dT) beads based on the presence of poly(A) tail. This procedure would also enrich for the transcript lacking a poly(A) tail but having a high adenine content [13, 27].

ribo-depletion Very common for general total RNA-seq analysis is the ribo-depletion. In this procedure, specific ribosomal sequences are used to target the most abundant rRNAs and deplete them from the sample [13, 27].

poly(A) depletion There are studies with focus on ncRNAs. Most ncRNAs lack a poly(A) tail; therefore, oligo(dT) beads can be used to deplete poly(A) transcripts [13, 15].

circRNA enrichment Due to the lack of free 5' or 3' ends circRNA molecules are theoretically resistant to exonucleases. RNase R is the most commonly used for such a procedure. RNase R degrades transcripts with free nucleotides at the 3' end. It has also been reported that RNase R treatment inexplicably depletes some circRNAs [13, 28].

Additonally, if the RNA biotype of interest is a small RNA, the aforementioned size selection procedure can separate RNA for a dedicated small RNA-seq.
It is important to note that, due to experimental limitations, neither procedure leads to perfect purification. All strategies merely enrich for an RNA biotype of interest.

2.3.2.2 RNA-seq library preparation

Due to the limitations of NGS technology, long RNA molecules need to be processed into shorter library fragments. Fragmentation is usually performed on the level of an RNA molecule and involves thermal treatment in the presence of metal ions. Alternatively, fragmentation can be performed on the cDNA level via a sonication procedure. For the latter, the RNA molecule needs to be reverse-transcribed into cDNA. Less common approaches are

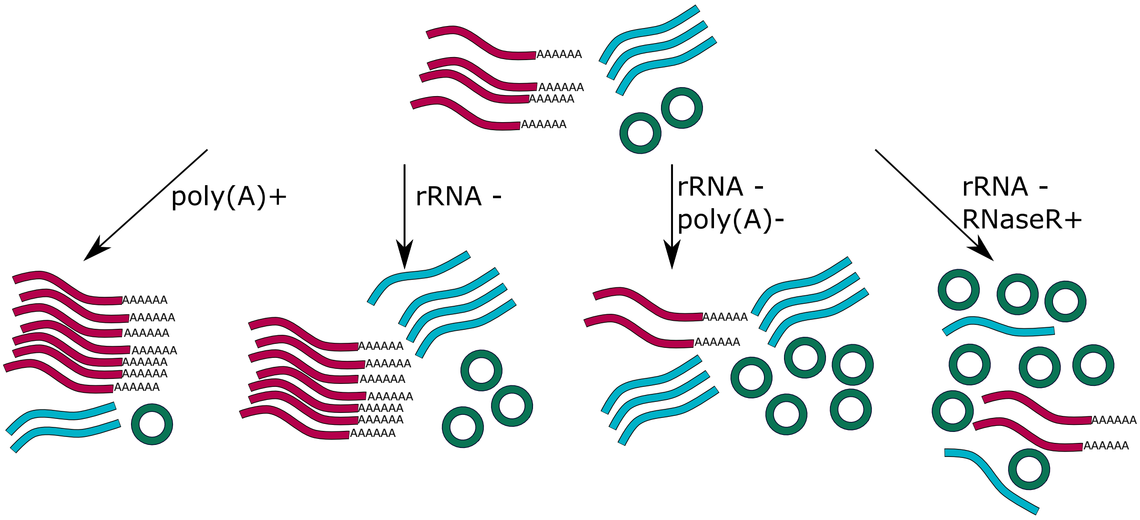


Figure 2.2: **RNA enrichment strategies** *On the figure we see the common RNA enrichment procedures with their specific marks: poly(A)+ – poly(A) enrichment; rRNA- – ribo-depletion; poly(A)- – poly(A) depleition; RNaseR+ – linear RNA depletion with RNase R treatment*

enzymatic or thermal treatments for fragmentation [29].
The reverse transcription requires a primer to start the process. There are three ways to provide a primers for reverse transcription [29]:

oligo(dT) primer If the library is poly(A)-enriched, a fairly straightforward approach is to use oligo(dT) primers that bind to the poly(A) tail

random primers A mix of random hexamers is added to the pool of RNA. They initiate transcription at random locations. In some protocols, random hexamers are combined with oligo(dT) primers to decrease biases in fragment generation.

pre-ligated oligo A pre-ligated oligo is added in some cases (e.g. small RNA-seq library preparation). This oligo provides a site for primer binding.

The nucleic acid fragmentation method incorporated in library preparation greatly influences the representation of circRNAs in the library. The size of the library fragment heavily influences the representation of circRNA in the RNA-seq library. Due to the fact that most circRNA size ranges within 200–400 nucleotides, libraries with the median fragment length size over that range are biased against shorter circles [13]. Furthermore, mild fragmentation conditions can cause failure to break the circRNA transcript and cause the rolling circle amplification product [30]. There is no clear record of how library preparation strategy involving cDNA fragmentation influences downstream analyses, but discussion and preliminary data suggest an unpredictable negative influence. The

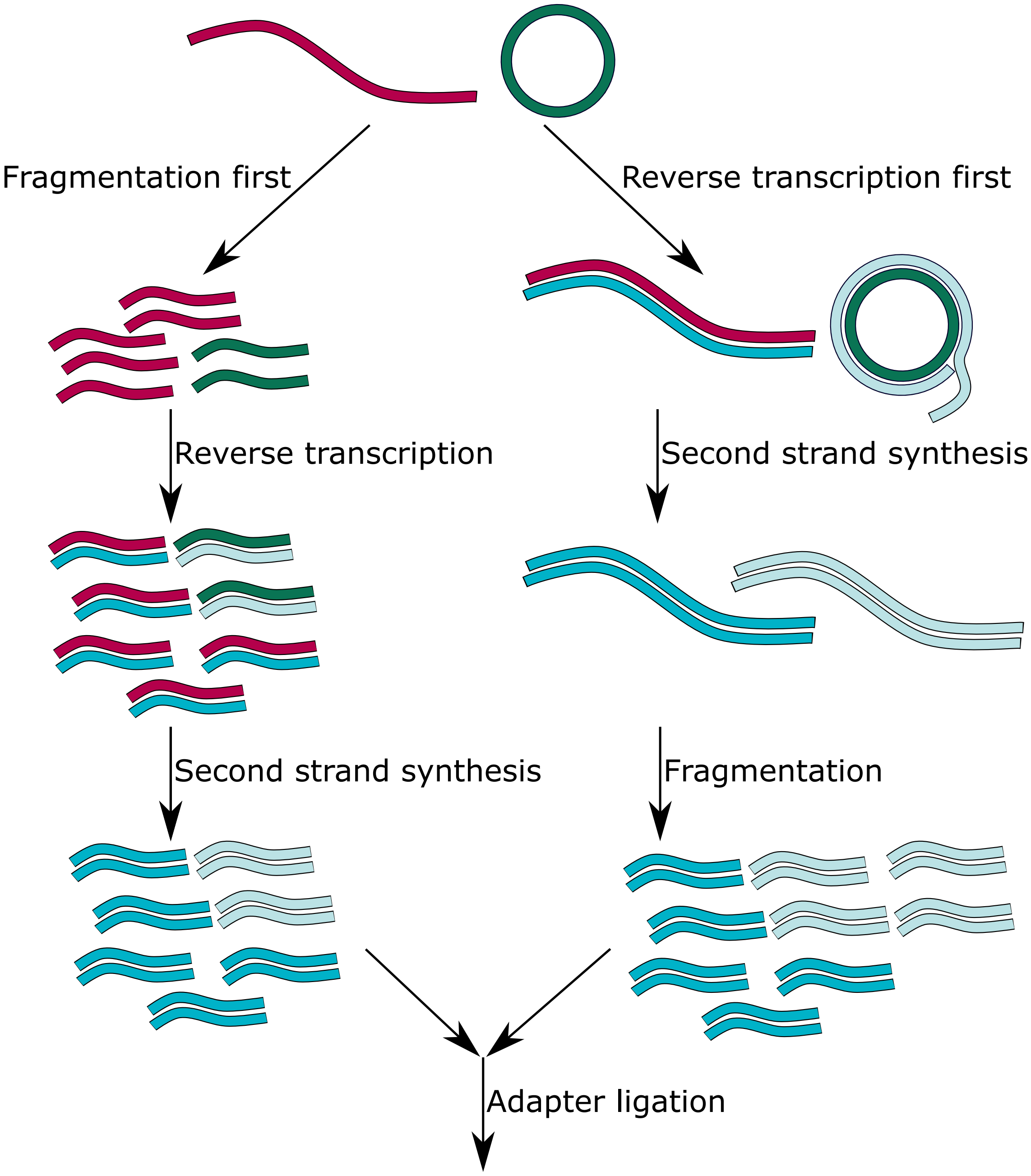


Figure 2.3: **Library preparation steps** *The two possible approaches divided by choice of RNA or cDNA fragmentation. While the effects are disregardable for linear RNA study, circRNA study is majorly affected. Rolling circle amplification has the potential to skew the number of generated fragments per molecule.*

final steps of library preparation are adapter ligation and PCR amplification. The purpose of adapter ligation is to provide the library fragment with the target sequences required by the sequencing platform. Adapter ligation is an inefficient process, therefore, PCR reactions that amplify only fragments with adapters on both ends are performed. The number of PCR cycles should be limited to a minimum because over-amplifying lowers the complexity of the library.

2.3.2.3 Sequencing procedure

The most commonly used sequencing platform is Illumina, which implements bridge amplification sequencing [31]. In this procedure, fragments are processed in a flowcell where they bind in nanowells to the sequence of the adapters. Once a fragment is attached to the flowcell it folds into a bridge-like shape by hybridisation of the second adapter. Multiple amplification reactions ensure a large number of clones corresponding to the forward and reverse sequence of the original fragment. The reverse strand segments are washed away from the flowcell in preparation for the next step.

The identification of the sequence of the fragment is performed as another polymerisation reaction in which the addition of a new fluorescently-tagged nucleotide is followed by the release of a fluorophore. The previous steps ensure a single unified signal for each base per spot in the flowcell [31]. Sequencing can be performed on one side of the fragment (single-end), or on both sides (paired-ed). It is rare that the entire fragment is sequenced. The part of the fragment that is sequenced is called a read.

2.3.3 Read mapping

There are two approaches to identifying the origin of each library fragment– de novo assembly and reference-based assembly. In de novo assembly, the reads are grouped based on their overlapping sequences into contigs [32]. These contigs serve as reference for mapping of the reads and later isoform assembly and quantification. The second option is to map the reads to a known reference sequence. This option allows for standardised analysis and limits error rates induced by wrong assembly. It is more common to use the genome as a reference sequence, but there are cases when users map to transcriptome sequences or even specifically selected sequences of interest. Although most reads map fully to a specific location of the reference genome, some reads can map partially to two or more separate locations. Such reads are called split reads. Based on computational restrictions of mapping to the genome, these reads are separated into two categories: splice junction reads and chimeric reads.

2.3.3.1 Splice junction reads

A processed RNA generally has all intronic sequences spliced out. Reads that span the splice sites are handled separately [33, 34]. For computational convenience, reads are marked as spliced only in the case of forward-splicing. In practise, this is done by staring the mapping of the read, introducing a skipped region in the mapping, and then mapping the rest of the sequence of the read. The process is guided by multiple hard thresholds that affect the length of the mapped parts of the read and the potential length of an intron.

2.3.3.2 Chimeric reads

When splicing occurs in back-splicing or trans-splicing fashion, reads are not reported as splice junction reads, but as chimeric reads [35, 36, 34]. Chimeric reads can also arise from fusion genes or be the result of a dedicated experimental procedure. It is a matter of post-processing to identify to which category the chimeric reads fall. In the case of circRNA, BSJs are identified by selecting chimeric reads with partial mappings in non-linear order according to the genome. The reads are then filtered based on the distance between the mappings [12, 37, 38] and sometimes on the basis of splice motifs [37].

2.3.4 Third-generation sequencing

2.3.4.1 Background

Naturally, the next level of technological advancement in sequencing is called third-generation sequencing. The term covers multiple platforms aiming to overcome the short fragment size that limits the NSG technology. Therefore, these approaches are often called long-read sequencing. The long-read technology solves major challenges in transcript reconstruction, particularly the identification of alternative isoforms and transcripts with integrated repetitive elements.

There are major platforms for long-read sequencing–PacBio and Oxford Nanopore. PacBio is the older technology, and in its most commonly used version, a single fragment is processed with polymerase that adds four distinguishable fluorescently labelled deoxyribonucleoside triphosphates (dNTPs) [39]. The growing DNA strand in a zero-mode waveguide nanostructure arrays provides optical observation volume confinement and enable parallel, simultaneous detection of thousands of single-molecule sequencing reactions [40]. The conjugation of fluorophores to the terminal phosphate moiety of the dNTPs allows the continuous observation of DNA synthesis over thousands of bases without steric hindrance [39]. Specialised adapters create a circular library fragment that allows up to 15 times the coverage of a molecule [39]. PacBio reports a median accuracy of 99.3%, without systematic error beyond fluorophore-dependent error rates [39].

Oxford Nanopore (often referred to as Nanopore) introduces a completely novel approach. The sequencing is performed by monitoring changes to an electrical current as nucleic acids pass through a protein nanopore [41]. Although 10-15 nucleotides pass through the channel at a time, only stretches of 9 nucleotides are the primary contributors to the current measurement in the pore. As the measurement is done on the basis of multiple nucleotides, the technology works optimally in sequences in which the nucleotide patterns have low similarity. Nanopore sequencing analysis leads to stretches of erroneously predicted nucleotides when a sequence with a simple repeat is encountered. Nanopore has two major advantages: RNA can be sequenced without time consuming multi-step processing; Nanopore sequencers are small, cheap and portable.

2.3.4.2 Long-read sequencing of circRNAs

Long-read sequencing of circRNA has to overcome multiple hurdles. The lack of distinguishing marker sequence (e.g. poly(A) tail), makes the specific enrichment of circRNAs problematic. Another problem is the lack of free 3'-end that is commonly used for ligation of target sequences for reverse transcription synthesis. The first long-read sequencing of circRNA was done with specific pre-selected primers diverging from the BSJs [42]. The products of the RT-PCR were sequenced with PacBio.

High-throughput sequencing of circRNAs was achieved in different independent projects, but all protocols have similar key steps [43, 44, 45]. RNase R is commonly used to deplete linear RNAs, however, its effect is limited by the presence of RNA structure at the 3'-end of the RNA molecule. To ensure a structure-free 3'-end, RNA extract is treated with poly(A)-tailing enzyme. This procedure greatly improves the efficiency of RNase R treatment. The reverse transcription is a step which all alternative protocols perform by rolling-circle replication of the circRNA molecules with random primers. The product of this reaction is DNA that contains the sequence of the template circRNA multiple times. The second strand synthesis requires a primer target. There are two approaches that solve this issue: poly(A) tailing and oligo(dT) primers or using adapters that facilitate the second strand synthesis. The cDNA fragments are then sequenced through Nanopore.

2.4 Common strategies in RNA-seq data analysis

2.4.1 Transcript assembly

Using NGS reads to assemble a reference of the source molecules is a challenging task. The problem is enhanced for RNA-seq assembly because the different isoforms of the same genes have overlapping regions. The key step towards the identification of the different isoforms is the detection of the splice junctions. Information from the reads that span the splice junction can sometimes be combined with the coverage of the exons is used to identify the differential features of the alternative isoforms [46]. Sequencing biases differ along the sequence of the gene, making the quantification of reads assigned to transcript features (exons, splice junctions, retained introns) difficult to normalise.

To avoid the issues with normalising exons, the assembly tools primarily focus on reads that span splice junctions [47, 48]. This includes split reads that are direct evidence of a splice junction occurring, as well as pair-end reads, in which pairs map to different exons. For the latter, it is important to know the fragment size of the library to predict whether the fragment can span multiple exons.

Such a strategy is enhanced when using a reference for the mapping and assembly of the reads [47, 48]. The sequence of the reference genome allows the tool to infer the sequence between the pair-ends of a fragment. Such use of pair-end reads shows the advantage of long library fragments and reads, as they are more likely to span a splice junction. However, longer fragments are produced with milder fragmentation procedures, which is problematic for the study of circRNA, since they fail to fragment under mild fragmentation conditions [30].

The first step of the reconstruction is shared by all tools – a splice graph is created based on the reads mapping to splice junctions. At the time of writing, there is no tool for performing assembly of linear and circular transcripts. By design, splice graphs are directed acyclic graphs and therefore are unable to hold information on BSJs. This forces the assembly of circular and linear RNAs to be performed separately.

There are two commonly used approaches to parse a splice graph for linear transcript assembly. One is that the approach used in Cufflinks [47] is maximum parsimony. Cufflinks uses an overlap graph, in which the sequenced fragments are nodes, and two nodes are connected if they overlap and have compatible splice patterns. The next step is to find the minimum set of transcripts that explains all reads in the graph. An alternative approach is used in StringTie [48]. StringTie iteratively extracts the heaviest path from a splice graph, constructs a flow network, computes the maximum flow to estimate abundance, and then updates the splice graph by removing reads that were assigned by the flow algorithm. This process repeats until all reads have been assigned.

Transcript assembly of circRNA transcripts is affected by an added challenge – the low abundance of circRNAs. Therefore, the assembly of circRNAs requires circRNA-enriched RNA-seq libraries. In some cases, assembly is performed only on the basis of reads that span the BSJ, which puts a threshold limit on the size of circRNAs that can be assembled [49]. Alternatively, tools for linear transcript assembly can be supplemented with additional scripts that ”guide” the output of linear assembly specifically to circRNA loci and subsequently reconstruct circRNA by ”lightweight” strategies [50, 51]. Both approaches struggle to robustly identify circRNA-specific transcript features and are very dependent on the available annotation [52].

2.4.2 Transcript abundance estimation

Downstream analyses require the estimation of the relative abundance of transcripts. The process can be separated into two parts– assigning reads to a reference and subsequent transcript quantification. Read alignment can be performed by dedicated mapping tool discussed in Section 2.3.3. However, such alignment is time-consuming and/or virtual memory-consuming. The availability of reference trascriptomes provides an lightweight solution to this problem. Sailfish processes the input transcriptome into k-mers (strings of particular size) organised into a hash table and suffix array. Those data structures allows for fast access to k-mers in a preprocessed (indexed) reference. When a read contains a particular k-mer, the location in the suffix array allows for a fast extension of the matching [53]. Kallisto also uses k-mer matching strategy, but in in this case the k-mers are organised in a directed graph of overlapping k-mers – De Bruijn graph. In this structure, k-mers are binned together based on k-mer compatibility and alternative splicing corresponds to different paths within the graph. A hash table keeps the information of the position of a k-mer within the graph and correspondingly to which isofroms it is assigned. The read is assigned to the minimal number of paths (isoforms) in the graph [54]. In Salmon, instead of a suffix array a Burrows–Wheeler transformed index is used – FM-index [55]. However, in practical terms, the algorithm still needs a fixed size of minimum acceptable length for a valid match equivalent to a k-mer size.

The quantification of transcripts is done predominantly by using an Expectation-Maximisation (EM) algorithm [56]. This approach uses iterative assigning of the reads per isoforms (E) and subsequent re-calculation of the assigned values (M). The re-calculated values are used as an input for the next (E) step and the procedure repeats until a threshold difference is reached. Alternatively, transcript quantification can be achieved as a by-product of the transcript assembly. StringTie calculates iteratively the abundance of each transcript as a maximum flow within a pre-selected section of the comprehensive splice graph. The calculated isoform abundance is then subtracted from the corresponding nodes of the graph, and the procedure is repeated until no more transcripts can be reconstructed. Transcript quantities are most commonly represented as Transcripts

Per Million (TPM) or Fragments Per Kilobase per Million reads mapped (FPKM) [57]. The key to those calculations is the transcript length, which allows for within-sample normalisation.

2.4.3 Differential expression analysis

The goal of differential expression analyses (DEA) is to identify genes, transcripts, or other features with significant differences between conditions. Most algorithms were designed for gene-level analysis, but can be used in a transcript-level study with slight modifications. There are also some strategies specifically dedicated to transcript-level studies [58]. The key problem in DEA is differentiating between a significant change in the expression levels of a feature and naturally occurring variance typical for biological systems. There are parametric [59, 60] and non-parametric [61] approaches toward DEA. Parametric approaches have become the general trend, as they can produce reliable results with a lower number of replicates.

The input for DEA is usually two pairwise sets of biological replicates. The replicates are needed to model the natural dispersion of the features. The variance is fitted into a Negative Binomial distribution to account for the fact that the gene dispersion can be much higher than the mean. With higher the number of replicates input into the analysis, it is more likely to correctly capture the natural variance of the genes. To this end, biological replicates are more valuable than technical replicates, as they account for more sources of variance. To make better use of a low number of replicates, the dispersion of all genes is fitted as a function between mean and standard deviation. This fit is used to adjust the dispersion [60].

For a proper calculation, it is important to normalise the counts between samples. The issue with using a simple normalisation by library size is that a few highly expressed genes will influence the calculations. Therefore, the common approach for between-sample normalisation is the use of quantiles as a reference point [62, 60]. The use of quantile normalisation works under the following assumptions: the number of up- and down-regulated genes is similar; most of the genes remain with unchanged expression levels between samples; differential and non-differential genes are equally subjected

to technical effects. As the most robust method is regarded the median ratios normalisation. A pseudo-reference sample is created based on the average of all samples. For each sample a size factor is calculated as the median of the ratios between gene counts in the sample and the pseudo-reference sample [60].
Transcipt-level studies require modified distributions to fit the alternative isoform levels. Cuffdiff2 [59] fits the counts in Beta Negative Binomial distribution while DRIMSeq [63] uses Dirichlet Multinomial mode. In both cases the read counts for a gene are fixed, and and those counts are distributed among the transcript within a gene for each sample.

2.4.4 Co-expression network analysis

Network analysis is a powerful tool to identify a group of transcripts or genes with similar functions or even the driving force behind their expression. The co-expresssion network analysis algorithms are clustering algorithms that make use of some properties of biological networks to improve clustering. The boom of development of co-expression network algorithms came with the popularisation of microarray data. Thus, the algorithms were developed for homoscadastistic data, assuming constant variance in measurement; see Section 2.3.2. The algorithms worked under a hierarchical model of gene networks, where nodes (representing genes) are connected to others in a hierarchical fashion, and those connection reflect functional association and control mechanism. Under this assumption, many control genes were identified as the hub (the most connected node) of a co-expression cluster. The most commonly used algorithm is Weighted Gene Co-expression Network Analysis (WGCNA) [64]. In this pipeline, first, a subset of the most varying genes is selected. The next step is to calculate the pairwise correlation of gene levels between samples using the Pearson correlation coefficient. Since genes are viewed as nodes in a network, the correlations is often referred to as similarity or adjacency. The correlation is transformed by a beta adjacency function (power). The parameter for this function is selected based on a so-called scale-free topology criterion, that is, the adjacency function parameter is optimised to maximise the number of hub nodes in the network. A topological overlap between genes is calculated, which ultimately recalculates the similarity between two genes based on their adjacency in the context of the other adjacency values of the network. A hierarchical tree is built based on the topological overlap values, and a tree-cut algorithm is used to separate the branches of this tree into clusters.

The assumptions valid for genes-level networks are not necessarily valid for transcript-level networks. For example, there is often one predominant isoform of a gene that corresponds to the majority of the gene expression level. Therefore, one cannot assume that a transcript-level network would have a scale-free topology. The core calculations in the algorithms are based on variance, and they cannot be directly applied to RNA-seq counts data that is heteroscedastic. Instead of fundamentally changing the algorithms, it is easier to transform the RNA-seq data with a variance stabilising function. A fast and straightforward method for this is the VST function of the DESeq2 package.

2.4.5 Gene set analysis

As a result of DEA or clustering, a set of genes or transcripts of interest is created. We need to identify patterns in their function or regulation mechanisms. For convenience, there are precompiled gene sets that allow fast hypothesis testing. The most commonly used gene sets are the Gene Ontology (GO) which contains information on the function of the genes, and KEGG which contains pathway information. Another useful source of gene sets is the Molecular Signatures Database (MSigDB), which contains curated sets for position, pathway, regulatory mechanism along with some set specific for particular biological problem. Any database that contains subsets of genes separated into categories can be transformed into gene sets suitable for testing. For the sake of testing, it is important to keep track of the overall set of genes at the start of the analysis called the gene universe. There are multiple techniques that can be used to find an association between genes of interest and a gene set. The combination of approaches used to be referred to as enrichment tests, but because of the prevalence of the Gene Set Enrichment Analysis (GSEA) package monopolised the name, simpler approaches adopted the names association or over-representation tests. However, to this day, the term ”enrichment” is used to indicate significant overlap with a gene set. Because the result of the analysis is usually functional association with a pathway, gene sets analysis is often referred to as pathway analysis.

2.4.5.1 Over-representation tests

The over-representation tests are a simple procedure that determines if a set of genes of interest gave a significant association with a known gene set. The analysis is divided into two parts. First, finding the overlap between the genes of interest and a gene set. Second, applying a statistical test to determine the significance of that overlap. The most commonly used tests are based on the hypergeometric, chi-square, or binomial distribution [65].

2.4.5.2 Enrichment tests

The hypothesis of enrichment test is that although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes can also have significant effects [65]. The most used example of such testing is provided by the GSEA package. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric. The goal of GSEA is to determine whether the members of the gene set are randomly distributed throughout the ranked list or primarily found at the top or bottom. The enrichment score is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov–Smirnov-like statistic [66].

2.4.5.3 Pathway Topology based tests

The previous methods consider only the number of genes in a pathway or gene co-expression to identify significant pathways, and ignore the additional information that genes are shared between biological pathways. Pathway topology (PT)-based methods have been developed to account about gene products that interact with each other in a given pathway, how they interact (e.g., activation, inhibition, etc.), and where they interact (e.g., cytoplasm, nucleus, etc.) [65].

2.5 RNA-RNA interaction detection

2.5.1 Background

2.5.1.1 Basics

RNA-RNA interactions (RRIs) play a key role in major functions from processing of nascent RNA transcript to regulation of their expression and localisation [5, 67, 68, 69, 70]. RRIs can influence nascent transcripts. Arguably, the most important RRIs is the splicing of nascent RNA molecules into a final transcript, guided by the snRNAs [5]. Other interesting examples of functional RRIs are snoRNAs. They edit nucleotides of RNA transcripts forcing restructuring of the molecule and allowing them to obtain their final functional conformation [71, 72]. Even tRNA codon recognition falls into the RNA-RNA interaction category [73]. RRI-based regulation of mature transcripts is not to be overlooked. By sheer abundance of examples, the most well-studied class of RRIs are miRNAs [67]. They can facilitate blocking RNA translation or decrease the abundance of target transcripts [67]. Functionally similar to eukaryotic miRNAs, bacteria sRNAs are responsible for the translation regulation of mRNA transcripts [68, 69]. Additionally, non-coding RNAs have been proposed to act as a transporter, allowing mRNAs lacking localisation signals to ”hitch-hike” to the correct cellular destination [70].

As opposed to protein-protein interactions that have been thoroughly studied, RRIs represent an untapped potential for new discoveries. One of the reasons for our surface knowledge of the functions of RNA is the challenges that studying RNA structure and RRI pose. Detection of RNA binding protein *in vivo* is in a stage where it can give reliable results [74]. As opposed to protein-RNA binding is guided by structural or sequence motifs [75] (binding methods that allows for sequence variations), RRI follow nucleotide level specificity, allowing for more precise regulation of biological processes. Due to the dynamic nature of the RNA structure experimental assays for specific RRI detection *in vitro* do not produce the same output usefulness as they do for proteins [76, 77].

2.5.1.2 RNA structure features

Similar to DNA, RNA has the potential to form helices (duplexes), based on complementary regions. Formation of RNA helices is based on base pairing nucleotides following standard Watson-Crick pairings (G-C, A-U) or participating in wobble pair interactions(G-U). Non-standard base pairings are possible, but they are facilitated by topological proximity due to a specific 3D structure. The formation of RNA helices is possible within the same RNA molecule (cis) or between 2 molecules (trans) . The combination of all cis-helices is commonly referred to as ”RNA structure”. Meanwhile, the term ”RNA-RNA interactions” is commonly used to indicate trans-helices. Stretches of the transcript sequence that are unpaired are called loops (long) or bulges(short). The approaches that lead towards RRI prediction are an extension of the principles that guide structure prediction. RNA structures often guide the formation of RRIs. Thus, to fully grasp the concept of RRI prediction, one first needs to understand RNA structure.

2.5.1.3 2D vs 3D

Naturally RNA structures are 3-dimensional. The more common approach towards RNA structure prediction is focussing on the 2D structure (secondary structure). Unlike proteins, where the 3D structure is essential for the prediction of binding sites, the secondary structure of RNA is often sufficient to predict functional interactions between RNA transcripts, due to the fact that it is defined by base pairing nucleotides.

2.5.1.4 RNA structure dynamics

As opposed to proteins, where one molecule often has one ’correct’ structure, RNA transcripts can have multiple alternative functional structures. The difference comes from the fact that while proteins have mechanisms that ensure the folding of the correct final structure, RNAs begin their folding as soon as the RNA molecule starts being transcribed. By the time the molecule is fully transcribed, the RNA transcript has undergone multiple changes in structure.

RNA structure can be key in the response of outside stimuli. The structure elements called riboswitches can change their structure

based on temperature [78], ligand binding [79] or pH change [80]. RNA structure can be view in two levels of abstraction: global and local. The global RNA structure focuses on the set of helices along the entire transcript sequence, while the local RNA structure focuses on a specific part of the sequence. Focus on the local structure allows us to gain a more adequate model of the functional significance of a particular region [81, 82].
The final RNA structures are the results of multiple events:

- RNA splicing – the splicing of nascent RNA happens co-transcriptionally, meaning that significant stretches of the RNA molecule are removed - leading to a change in structure [83]
- Protein binding – proteins can bind to RNA (e.g. QKI, Hfq) and force a change in the structure [69, 84, 85]
- RNA editing - the process of chemically changing a base forces an alteration in the RNA structure [86]
- Speed of transcription - the speed of transcription is decisive as to whether or not any of the aforementioned mechanisms of action are given time to implement [87]
- trans RNA interactions - RRIs can cause a functional change in RNA structureby blocking a site of helix formation [88]

2.5.1.5 Challenges of RRI detection

Computation prediction of RRIs in the transcriptome is achieved by pair-wise testing of all RNA transcripts. Putative RRIs are prone to a high false discovery rate (FDR). Even if the interaction pair is predicted correctly, the exact base pairs could be wrong. There are *in vitro* experimental methods for RRI detection [77, 76, 89] However, the nature of those methods does not provide nucleotide-level information. Therefore, the identified pairs from the experimental assays need to be processed by computational tools to identify the precise base pairs in interactions. The combination of experimental data complementing computation tools allows for a significant decrease of the search space of potential RNA transcript pairs, leading to a significant decrease of the FDR. Nonetheless, even a combination of *in vitro* experiments and computational predictions fails to predict *in vivo* RRIs, due to the disregard of the structural dynamics of RNA.

Potential trans RNA interaction loci can be blocked by RNA structure. This type of blocking is represented by the concept of accessibility. The structure of both participating RNAs should allow the trans interaction to occur. Computationally accessibility can be used to narrow down the search space of potential trans RRI by applying an RNA accessibility factor (see section Computational approaches). However, prediction of RNA accessibility as a purely computational task is flawed due to the fact that RNA structure can be reshaped via all the events mentioned before.

To reiterate, the interaction of RNA with other proteins and RNA, as well as the possible alternative splicing events, makes a purely computational prediction of RRIs unreliable. The theoretical complementarity between the two RNA molecules is not sufficient for a functional RRI prediction. However, a plethora of *in vivo* experimental methods have been developed to improve our chances of understanding the RNA interactome, by significantly limiting the search space for potential helices to 20-80 nucleotides away from the interaction spot (see section Experimentally driven strategies).

2.5.2 Computational strategies for RRI prediction

2.5.2.1 Minimum free energy approach

The intuitive approach to computational prediction RNA structureand RRI is to try to maximise the number of base pairs within a molecule pair complex. It is important to understand that the biochemical purpose of base pairing is to minimise the free energy of the complex. We can optimise the base pairing algorithm by enhancing it with experimental information on the energy gain of a particular base pairing, giving a thermodynamic aspect to the folding strategy. Systematic experiments were conducted using a series of oligonucleotides to calculate the energy gain at different temperatures, and the parameters for a nearest neighbour thermodynamic model were developed based on those calculations [90]. There are other aspects that influence the RNA structure: the stacking interaction of paired nucleotides (stabilising), and the dangling energies of free nucleotides (disruptive). Over time, the thermodynamic models have been enhanced by recalculations based on available structural information, but given the small number of known RNA structures, the adjustments have limited scope [91].

2.5.2.2 Approaches that utilise evolutionary information

Thermodynamic folding decreases in reliability with an increase in the length of the input sequence. As we have already established, knowledge of the RNA structure is needed to correctly predict RRIs. There is, however, an alternative approach to limit the helix search space. Functionally relevant RNA helices are conserved through evolution. Evolutionary information can be viewed in two aspects: positive and negative [92, 93]. Positive examples are mutations occurring in a functionally relevant RNA helix that are compensated (often by another mutation) in order to salvage the structure. Retention of a potential helix sites through evolution provides positive information that a helix is functionally relevant. Unpaired regions often allow mutations without functional loss [94, 93], thus providing negative information for the location of loops and bulges. Positive information can be separated into two types of compensatory events: covarying bases at the same location in the sequence [93]; addition of bases at the edges of the helix [94]. Note that base pair change can occur as one-sided variance (often creating a wobble pair) or full covariance (both bases are changed).

For the occurrence of two covarying bases, two sections of the genome need to mutate in compensatory fashion at the same time. When searching for compensatory events indicating RRI, we need to keep in mind that one RNA often interacts with multiple partners. The simultaneous equivalent mutation of all interaction partners is very unlikely. In conclusion, covarying bases are a strong indicator for RRI when ONLY two genomic loci are involved. While, for trans interactions that need to be consistent between multiple loci, it is more common for the region to be simply conserved or with rare instances of one-sided variation [95].

An example of positive evolutionary information is the change in nucleotide usage that preserves the ability of the two bases to form the RNA helix. If the structural or interacting RNA element has an important function, then observable mutations need to compensate for each other to salvage the structure. Thus, the retention of potential helix sites through evolution provides positive information that a helix is functionally relevant. Opposite to the base-paired regions, the unpaired regions often allow for mutations without functional loss [94, 93], thus providing negative information for the location of loops and bulges.

The positive information can be separated into two types of compensatory mutation events: covarying bases on the same location in the sequence [93]; addition of bases at the edges of the helix [94]. The base pair change can occur as either one-sided variance (often creating a wobble pair) or full covariance (both bases are changed). For the occurrence of two covarying bases, two sections of the genome need to mutate in compensatory fashion at the same time.

When looking for compensatory events that indicate RRI, it should be noted that the RNA molecule often interacts with multiple partners. Simultaneous compensatory mutations of all interaction partners are very unlikely. While, for trans interactions that need to be consistent between multiple loci, it is more common for the region to be simply conserved or to have rare instances of one-sided variation [95]. In conclusion, co-varying bases are a strong indicator for RRI when ONLY two genomic loci are involved.

2.5.3 Experimental strategies for RRI detection

2.5.3.1 Transcript-specific pull-down approaches

The most straightforward approach to the experimental search of interacting pairs is the pull-down assay. The need for *in vivo* detection of interactions requires focus on a specific transcript of interest. In MAPS [96, 97] the RNA transcript of interest is tagged with an aptamer and introduced into the cell with an inducible plasmid. The cell extract, which contains the tagged transcript and its targets, is filtered through a column with the MS2 protein that recognises the tag. The filtered extract sequenced as RNA-seq and compared to control library produced without the selection tag. Comparison can be performed using any differential expression pipeline (e.g. DESeq2 [60]). The downside of this method is the identification of not only direct targets but also secondary targets (see Figure 2.4).

A similar approach is provided by RAP and RIA-seq [98, 99], where the RNA of interest is pulled down by specific biotinilated probes. RAP is enhanced by AMT cross-link - a psoralen derivative (see

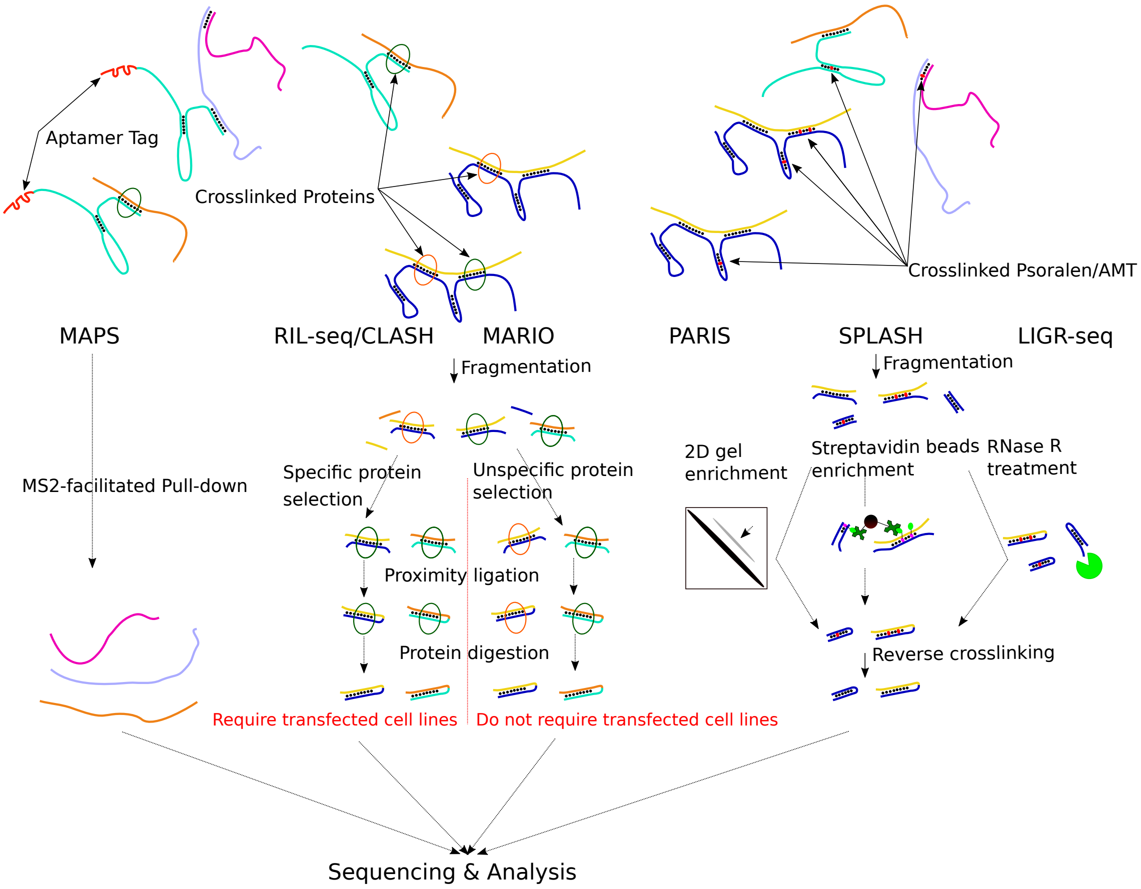


Figure 2.4: *Overview of several experimental methods that assess the presence of RRIs* in vivo*. MAPS stands out from rest of the methods as a straightforward pull-down experiment. The rest of the methods follow five common keys steps: 1) cross-linking; 2) enrichment; 3) fragmentation; 4) proximity ligation; and 5) removal of the cross-link. All resulting RNA fragments are processed as transcriptome libraries.*

Section 2.5.3.3). All of the pull-down approaches need to be supplemented by computational RRI identification to reach nucleotide resolution and filter false positives.

2.5.3.2 Protein-mediated approaches

CLIP-seq is a unified name of a set of techniques for detection of RRI via RNA proximity ligation. The key feature of CLIP-seq protocols is the enrichment of specific duplexes by protein pull-down. Again, as a specific case, the miRNAs allow for *in vivo* interaction prediction with minimal data. Due to the fact that the function of miRNAs depends on the AGO complex, CLIP-seq data (and its derivatives) can be used to identify miRNA-mRNA interaction pairs [74, 100]. CLIP data is based on cross-link of RNA-protein complexes, subsequent protein purification and degradation, and sequencing of the selected RNA stretches. The resolution of CLIP data allows narrowing down the search space of potential miRNA target regions, while simultaneously providing *in vivo* specific interactions information.

There are other RRIs that are facilitated by protein binding, but the participating transcripts are longer than miRNAs. There are techniques that focus on using those proteins as anchors that keep the RNA molecules in close proximity to ensure ligation between the two RNA molecules (Figure 2.4). Firstly, cross-linking RNA and tagged proteins, subsequently selective immunoprecipitation of those proteins is performed to enrich the interaction complex. RNA is fragmented, and proximal RNA fragments are ligated. Finally, the protein is degraded to leave a chimeric RNA fragment, which, after sequencing, provides information of the two interacting RNAs. This strategy is applied in RIL-seq [101] and CLASH [102]. However, limiting the study of RRI to a single protein is rarely the aim of interactome studies. Recently, MARIO [103] was introduced that applies the same protein anchor strategy, but on a full proteome scale. In MARIO, protein selection is done by tagging the protein with biotin. Another advantage of MARIO is the presence of a biotinated linker region added to the chimeric fragments, allowing selective enrichment of the chimeras. Very recently, an alternative to MARIO strategy has been presented. In RIC-seq [104], the enrichment method is MNase treatment, which fragments RNA not bound to proteins. During the proximity ligation, a single tagged nucleotide is used as a tagged linker.

2.5.3.3 Psoralen facilitated approaches

The dependence on protein binding for the identification of RRI is undesirable. To circumvent this, a series of techniques reliant on the compound psoralen were developed. Psoralen and its derivatives can intercalate within an RNA duplex and upon irradiation with 365 nm UV be cross-linked to adjacent pyrimidines on opposite strands of the duplexes [105]. The importance of psoralen is the ability to reverse the cross-link at 254 nm UV. Pairing RNA transcripts are processed in different ways, depending on the protocol. The end goal of all strategies is a chimeric fragment, produced by proximity ligation of the RNA transcripts that forms a duplex.

In Figure 2.4 you can see the similarities of the experimental methods. The primary differences between psoralen-based protocols stem from the duplex enrichment strategies. In PARIS [106], the duplexes are manually cut from 2D gel electrophoresis. In SPLASH [107], the psoralen derivative is tagged with biotin for streptavidin pull-down. More special is the case of LIGR-seq [108], where enrichment is done by exonuclease (RNase R) treatment. Due to the nature of the treatment, the LIGR-seq protocol produces a lot of misleading ligations. Thus, the protocol requires control libraries, and the corresponding bioinformatics pipeline, is based around a series of statistical tests to determine enrichment of chimeric fragments between control and treatment libraries.

The selection of chimeric fragments can be enhanced by pull-down of sequences of interest, an idea applied in COMRADES [109]. They also use azide-modified psoralen for pull-down to avoid the limited permeability of the cell to biotinated psoralen [107].

2.5.3.4 Computational processing of duplexes

The goal of the detection of duplexes based on chimeric fragments is to decrease the search space of RRI to sequence stretches of rarely more than 100 nucleotides. There is quite some variance between the bioinformatics pipelines proposed together with the original experimental protocols. However, every pipeline can be summarised into a few key steps. We also propose common tools that can be applied in the following steps:

1. **Chimeric fragment detection** Mapping the data with a tool that can handle split read mapping. Good choices are BWA-MEM [35] or STAR [34].

2. **Filter non-chimeric fragments** The results should be filtered for known splice junctions. Requires simple script for STAR output, one can use the pipeline from PARIS, while for BWA-MEM - SPLASH.

3. **Binning the reads** The overlapping reads from chimeric fragments corresponding to the same RRI need to be collapsed into a single region. Common tools for such a task are BEDTOOLS [110] and the GENOMICALIGNMENTS R package [111]. Those tools can also be used to annotate the binned regions.

4. **RRI detection** This step is sometimes skipped, but it is important to provide nucleotide resolution to the RRI. Any RRI detection tool should be able to handle such short sequence stretches. In [108] the authors use RactIP [112]. For the psoralen-driven strategies quick false discovery filter can be applied, to check for adjacent pyrimidines on the opposite strands.

Split mapping requires both parts of the reads to have matching sequence to the reference (usually at least 15-20 nt). Usually, reads are mapped to the full genome reference. In that case, if the read spans simultaneously splice junction and a chimeric junction, then it is required to have multiple sufficiently long matching regions. With fragments of about 60-70 nt length that is very unlikely. A solution to this problem is using the transcriptome as a reference to remove the need for splice junction mapping seeds. To avoid multi-mapping to different isoforms of the same gene, usually the longest isoform of the gene is selected. The inherit downside of the transcriptome mapping is the loss of information of RRIs in the intronic regions of the genes.

Chapter 3

A new computational method for circRNA assembly – CYCLeR

The work presented in the scope of this chapter was performed by the doctoral candidate, under the oversight of prof. Irmtraud Meyer (MDC-Berlin). I researched the field and identified the need for a novel tool. I selected the strategy and algorithm that the new tool would employ. Subsequently, I developed the tool along with another novel tool designed for circRNA RNA-seq simulation. I determined the nature of the benchmarks and conducted all analyses. The work described in this chapter is published in the journal *Nucleic Acids Research* under the title *CYCLER–a novel tool for the full isoform assembly and quantification of circRNAs* [52]. All materials used in the publication are available under CC BY-NC. Most figures, tables and captions presented in this chapter are reproduced from the publication. The final text of the publication is the result of multiple rounds of comments and guidance from Prof. Irmtraud Meyer. The GitHub page of the tool was improved thanks to the feedback of Dr. Altuna Akalin (MDC-Berlin).
Note: The use of ”we” throughout the text refers to the author-reader collective.

There is a plethora of tools used for circRNA research with distinct aims and approaches [113, 114, 115, 116, 30, 117, 118, 49]. None of those methods, however, cater for the need to identify the full sequence of all circRNAs, as well as the simultaneous estimation of the expression level of linear and circular RNAs. This is why I developed the novel tool *CYCLeR* (Co-estimate Your Circular and Linear RNAs). For the assembly of full-length circRNA isoforms, *CYCLeR* employs a comparison of two types of RNA-seq data: total ribo-depleted (control) and circle-enriched. The initial step of the algorithm involves identifying circle-specific features. Subsequently, it utilizes a flow-based algorithm to predict the most likely set of circular RNA(circRNA) alternative isoforms. These circRNA transcripts are then transformed into a pseudo-linear isoform profile, enabling the estimation of the abundance of both linear and circular transcripts through an expectation-maximisation (EM) approach. [52]. In this chapter, I describe the main features of the *CYCLeR* algorithm and prove its advantage over alternative strategies by means of an extensive benchmark.

3.1 Introduction and motivation

3.1.1 Assembly challenge

Both linear RNA and circular RNA are a product of the splicing of a nascent RNA transcript [9]. The mainstream use of Illumina sequencing has led to great strides in the identification of novel RNA isoforms. However, RNA-seq reads originating from linear and circular splicing isoforms cannot be easily assigned to their transcript of origin. This makes transcript assembly challenging, see Fig. 3.1. An additional issue is estimating the relative abundance of circular and linear RNA. Direct evidence of circRNAs from RNA-seq data can be derived solely by the detection of reads that map to the location of the circular splicing – the so called back-splice junction (BSJ), see Section 2.2.2. However, BSJ-spanning reads provide only limited information on the complete sequence of circRNAs, which hinders the assembly and qualification of circRNA isoforms.

Common transcriptomic analysis, such as differential expression and co-expression correlation analyses, as well as RNA structure and interaction identification, all rely on prior knowledge of the transcript sequence and relative abundance. The widely utilized transcriptome assembly tools [47, 48] operate by constructing a directed acyclic splice graph. In this graph, nodes and edges are established based on the mapping of forward junction-spanning reads

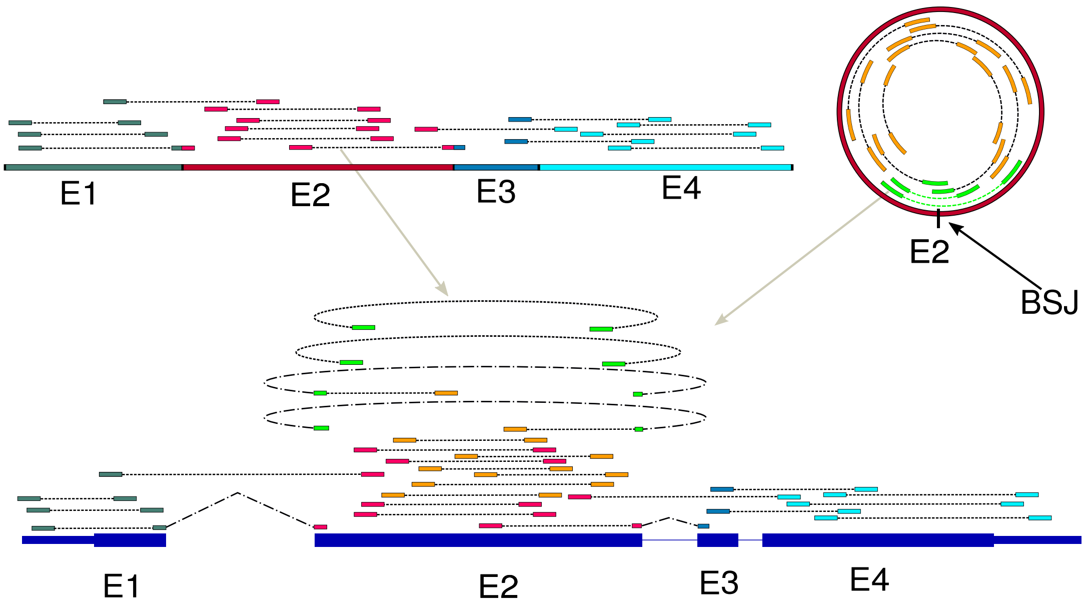


Figure 3.1:* Challenge of identifying circRNAs from RNA-seq data.** *Typical, raw transcriptome data from linear and circular splicing isoforms (top left and right) comprises a multitude of pair-end reads covering the exons of these isoforms (E1 etc, colouring of pair-end reads according to the exon from which they derive). In order to infer the original splicing products from these raw transcriptome reads, they are typically first mapped to the genome (bottom). Most of the mapped reads will not cover splice sites (exon-intron boundaries) and could either derive from a linear and circular splicing isoform. One challenge is that only reads spanning a* back-splice junction *provide direct evidence for circRNAs (marked in light green). As is also clear from this picture, the correct identification and quantification of circRNAs cannot be achieved without the simultaneous identification and quantification of the linear splicing isoforms. Thus, if the linear splicing isoforms of a gene are known up-front, their correct quantification needs to be estimated in conjunction with the identification and correct quantification of unknown circRNAs.* ***Reproduced from Stefanov et al. (2022)

in the transcriptome data. The assembly algorithms subsequently utilize these splice graphs as a base for the reconstruction procedures; details in section 2.4.1. Although the general idea is valuable, significant adjustments are necessary to accommodate the cyclic splice graphs essential for the assembly of circRNAs [52].

3.1.2 circRNA enrichment

Robust transcript assembly requires adequate read coverage across all exons and splice junctions. The presence of linear splicing isoforms originating from the same host gene can considerably bias the assembly of circular transcripts. The low relative abundance of circular RNA compared to linear [13] poses a challenge that can only be resolved by additional experimental procedures; see Section 2.3.2.1. To facilitate the study of circRNA, it is common to use transcriptome libraries, specifically enriched for circRNA [13]. Throughout the text, RNA-seq library generated by any method for circular enrichment are referred to as *circRNA enriched library.*

3.1.3 circRNA tools classification

To conduct a fair benchmark of *CYCLeR* in comparison to existing tools for the identification and quantification of circular RNAs (circRNA), I begin by categorizing the existing methods based on their objectives [52].

The tools most commonly used in circRNA studies, and naturally the representatives of class I, are known as CircRNA identification tools. Class I tools identify BSJ-spanning reads in RNA-seq data. The quantity of the reads mapped to the same BSJ is the measure that class I tools provide for the quantity of circRNAs. A limitation of these tools is their inability to ascertain the complete sequence of circRNAs.Furthermore, they lack the ability to detect AS of transcripts sharing the same BSJ site. However, these tools provide an important first step in any circRNA study. The lists of BSJs provided by tools from class I serve as an input of the tools of the other classes.

The objective of the tools from class II is to detect circRNA AS events and match them to a BSJ site. Class II tools sue the predictions of class I tools as input. The AS of linear RNAs is an obstacle to the detection of AS events of circRNA molecules. Consequently, class II tools require circRNA enriched libraries for optimal performance. The strategies employed in class II tools can be divided into two subgroups. The first approach relies on utilizing mate-pair information from paired-end reads spanning a BSJ for AS detection [30, 115]. However, a drawback of this method is its restricted sensitivity to AS events, limited by the insert size of the RNA-seq library. The second approach is utilized in the CIRCexplorer2 [50] pipeline and shares a strategy similar to *CYCLeR*. It involves the use of circle-enriched and total ribo-depleted libraries to identify forward splice junctions of circRNAs. An advantage of this strategy is its ability to overcome the limitations associated with the library insert size.

Reads spanning a BSJ typically constitute only a small fraction, approximately 0.1%, of the entire library. Depending exclusively on BSJ-spanning reads for quantification is unreliable due to their limited representation [52]. As mentioned earlier, the quantification provided by tools from Class I and II is lackluster and requires improvement. Two distinct strategies have been developed to tackle this problem. Due to their shared goal, I categorize them as class IIIa and IIIb.

Sub-class IIIa includes tools that estimate circRNA abundance as a ratio between back-splice junction (BSJ) and forward-splice junction (FSJ) counts [118, 119]. Both representatives of this subclass employ heuristics to estimate circRNA levels based on linear RNA abundance [118, 119]. Although the reported results align with qPCR benchmarks [118, 119], these tools do not enable the deduction of relative expression levels for all alternative linear and circular isoforms.

A completely different approach is employed by the single member of the sub-class IIIb – sailfish-cir [116]. In the subsequent step, sailfish-cir utilizes the created circRNA models and the provided linear annotation to perform simultaneous quantification of both linear and putative circular transcripts.

Recently, the first tools aiming to recover the full sequence of a circular transcript have been published. Due to their similar strategy and limitations, they are assigned to class IV [49, 51]. The initial stage in the full sequence assembly of circRNAs invariably involves BSJ detection. Even with libraries that have undergone circRNA enrichment, the BSJ output from tools belonging to class I or class II is indispensable. This is because linear RNA transcripts may still be present in the samples, potentially disrupting circRNA assembly. Additionally, the detection of the BSJ site resolves a common challenge in linear transcript assembly - identification of start and end exons of the transcript [52]. The assembly strategy applied in CIRI-full employs a similar approach to class II tools, where mate-pair information is used to predict the full length of the circRNAisoforms. This method is effective only in cases where the full length of the circle is covered by the paired-end BSJ-spanning reads. Naturally, this approach has the same limitation as the ones discussed in class II. This insert size bottleneck allows only for the identification of circular isoforms of up to 600 nt length, even when

using the recommended input of 2x250 nt paired-end RNA-seq libraries [49]. The algorithm of CircAST works by finding the minimal set of predicted isoforms that accounts for all FSJs [51]. In the output of class IV tools, the alternative circRNA splicing isoforms sharing a BSJ site are quantified as fractions of the total reads mapping to a BSJ site [52].
All tools across different classes that aim to identify novel circular RNA splicing events benefit from a circRNA-enriched library as input [52]. Class V is dedicated solely to the tool presented here–*CYCLeR*, as it provides functionalities that other tools lack. Particularly, a thorough transcript assembly algorithm and simultaneous quantification of novel circRNA and linear transcripts. A crucial distinction in data requirements compared to other tools is that *CYCLeR* mandates the inclusion of replicates. The summary of the differences can be easily tracked in table 3.1.

As an alternative to all strategies that rely on RNA-seq data, the members of class VI use Nanopore sequencing. The product of a rolling circle amplification of the reverse transcription of circRNA is sequenced by the Nanopore, see Section 2.3.4.2. The data from the long-reads is then processed and collapsed into a set of circRNA isoforms [44, 43]. A unifying trait of the experimental procedures required by the tools of class VI is high workload requirements. Both linear depletion and high replicate number are required for a reliable analysis [44, 43].

3.2 Simulation of RNA-seq data

3.2.1 Simulated transcript generation

At the time of writing, there is no comprehensive gold standard dataset available for benchmarking AS of circRNAs. The only option for a reliable benchmark in this study is to simulate a circRNA set to serve as a reference for RNA-seq simulation. My primary goal when designing the reference set is to mimic the key features of real data. To achieve this, I utilize publicly available *D.* melanogaster head data with available RNase R treatment to identify authentic BSJ sites and genomic characteristics specific to circular RNA (circRNA). This approach enables the detection of novel genomic features such

Class	Common reference name	Practical purpose	CE*	Representatives
Class I	CircRNA identification tools	BSJ Identification and quantification	no	CIRI2, CIRCexplorer KNIFE, etc.
Class II	CircRNA characterisation tools	circRNA AS event identification	yes	CIRCexplorer2, CIRI-AS, FUCHS
Class IIIa	CircRNA quantification tools	Improved circRNA quantification, based on BSJ to FSJ ratios	yes	CLEAR, CIRIquant
Class IIIb	CircRNA quantification tools	Improved circRNA quantification by using model-based framework	no	sailfish-cir
Class IV	Tools for full-length assembly of circRNAs	Full-length assembly of CircRNAs and relative CircRNA isoform abundances	yes	CIRI-full, CircAST
Class V	-	Full-length assembly of circRNAs and simultaneous linear and circular RNA abundance estimation	yes	*CYCLeR*
Class VI	-	Full-length assembly of circRNAs based on specifically generated Nanopore library	yes	CIRI-long, isoCirc

Table 3.1: **Classification of existing methods for circRNA identification and quantification.** *according to their goals and the input they require. Tools in column CE (Circle Enriched) denoted by a 'yes' require circRNA enriched libraries as input for optimal performance.* ***Reproduced from Stefanov et al. (2022)***

as exons, junctions, and retained introns, which are then incorporated into the simulated dataset. Figure 3.2 illustrates the necessary features that a simulated dataset should possess. A secondary goal when designing the data set is to ensure the high complexity of the transcript assembly task to test the potential of the tools to discover novel isoforms in challenging conditions. To this end, my simulated set needs to include cases of overlapping circular as well as linear transcripts. Table 3.2 contains a summary of key design characteristics of our set of simulated transcripts.

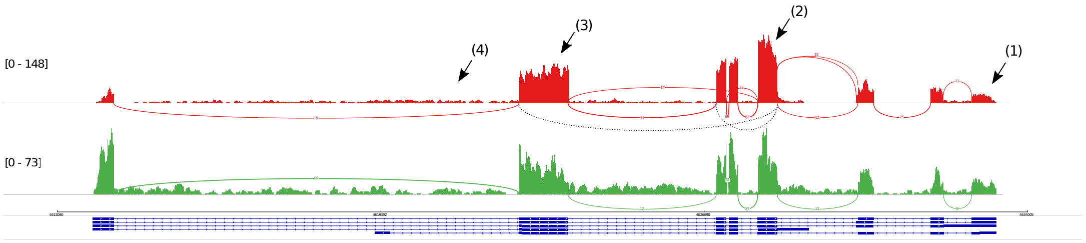


Figure 3.2: **Simulation of the location of the RNA sequence data for the 5-HT2A gene.** *Sashimi plot comparing simulated (red) versus real (green) RNA-seq data. BSJs are marked with dotted line. The necessary behaviour to consider a simulated sample "realistic": (1) Decrease in coverage around the start/end of linear transcripts, (2) relative decrease in coverage around back-splicing sites, (3) CG-bias in coverage and (4) intronic "noise" caused by unspliced transcripts.* ***Reproduced from Stefanov et al. (2022)***

It is essential for the data set to contain junctions that are not present in available annotations. Therefore, I map *D.* melanogaster adult head RNA-seq data [9] with STAR [34] and retrieve the full set of junctions in the real data. Subsequently, the alignment files are processed with the *SGSeq* R package [120] to create a comprehensive list of exons and junctions. I selected exons within BSJ loci, based on the unified results from CIRCexplorer2 [50] and *CIRI2* [117]. I filtered the exons and junctions based on enrichment after RNase R treatment. I removed monoexonic and diexonic circRNAs from the final reference as they do not pose an algorithmic challenge.

To generate the set of alternative isoforms, I employed a selection process for internal exons associated with each set of edge exons. This selection involved a Bernoulli trial with a success rate of 0.75. I repeated this selection process a total of (number of internal exons)/3 times. Naturally, only the unique transcripts were chosen for the following simulation steps. Additionally, I included all annotated linear transcripts expressed from the same genes as the circRNAs to the dataset.

An important part of the RNA-seq processing is normalisation of feature abundances. To test normalisation procedures I needed to provide simulated libraries with a realistic set and abundance of transcripts. The initial selection of transcripts, selected for simulations, was supplemented with an additional 10 000 randomly chosen protein-coding transcripts, which serve as placeholders for transcripts depleted during circRNA enrichment procedures. In addition, I incorporated 5,000 randomly selected non-coding transcripts into the dataset. These non-coding transcripts symbolize the linear transcripts that experience an increase as an outcome of circRNA enrichment procedures. The final library size was determined to align with a common sequencing library depth, which typically consists of around 25-40 million reads. [52].

Reconstruction problem	Dataset design
Identifying circRNA exons	Selected exons after circRNA enrichment
Identifying un-annotated exons	Integrated novel SJ from STAR output
Overlapping linear *AS*	Included overlapping linear transcripts
Overlapping circRNA *AS*	Included overlapping circular transcripts
Nascent RNA noise	Included full gene sequence

Table 3.2: **Benchmarking set design goals.** *The benchmarking dataset is designed to test the capability of different circRNA transcript reconstruction tools to deal with common problems for circRNA reconstruction-summarized in the table.* ***Reproduced from Stefanov et al. (2022)***

To designate transcript quantities, I used random sampling, with specific ranges selected empirically. I selected the following ranges for linear and circular transcripts respectively - from 10 to 40 and from 8 to 20. In the case of nascent RNA simulations, I used a factor within the range of 1-1.5. I assigned quantities on the lower side of the ranges to transcripts that contain unannotated exons and retained introns. In accordance with empirical observation of RNA-seq data, I adjusted the quantities of the reads from the simulated circular transcripts to amount to approximately 0.1% of the simulated linear reads. The calculation of the number of simulated reads per transcript involves the length of the transcript multiplied by a factor of 50, which has been determined empirically. The selection of this factor aims to replicate the typical coverage of a gene within a library. [52]

$$\# \, of \; simulated \, reads \, per \; transcript = \frac{Length \; of \; transcript \cdot Factor}{50}$$

Reads corresponding to linear RNA simulation are produced with polyester [121], while reads derived from circRNAs are simulated with polyestercirc; see section 3.2.3.

3.2.2 Reference sets

Based on the results from dedicated circRNA long-read studies [44, 43], we can make assumptions of the frequency of circRNA *AS*. In accordance to those studies, I simulate as most frequent *AS* event alternative circularisation (alternative BSJ occurring from the same gene), at lower rate exon skipping and alternative 5'/3'-splicing, and sporadic occurrence of intron retention.

The goal of the benchmark is to showcase the performance of the tools in common cases of circRNA assembly, as well as how they handle harder cases that rarely occur. Thus, I have separated the simulated transcripts into two reference data sets, Figure 3.3. Common cases of AS in circRNA involve a singular AS per locus. The data set focusing on these common cases is referred to as a reference set. The more complex set that involves extreme cases that challenge the assembly is referred to as *high complexity* reference set. A summary of the characteristics of the reference sets is available in Table 3.3.

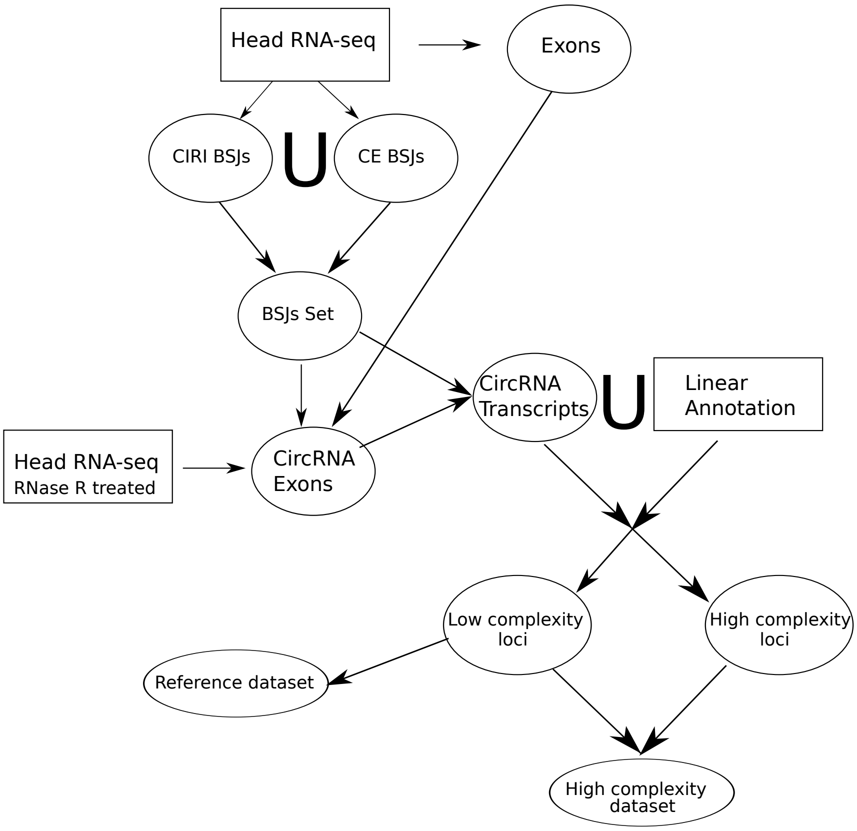


Figure 3.3: **Selection of reference sets.** *The simulated dataset contains cases that present common challenges for circRNA assembly that we know to be present in real data (see section "Reference sets") as well as even more challenging cases. Iselect a set of lower complexity cases to better highlight the differences in performance of the tools and employ the higher complexity set to investigate the limits of transcriptome assembly based on RNA-seq data. Our* reference set *contains loci with a low number of overlapping* AS *events. Our* high complexity set *expands the* reference set *by combining it with circRNA transcripts with multiple overlapping* AS *events. The symbol "U" indicates merger of sets.*

Dataset	Reference set	High complexity set
Unannotated exons	Yes	Yes
Retained introns	Yes	Yes
Overlapping linear RNA *AS* events	Yes	Yes
# of overlapping circRNA *AS* events	≤ 1	All

Table 3.3: **Reference vs High complexity sets** *This table specifies the differences between the* High complexity set *and the* Reference set*. Both datasets allow us to assess the evaluation of the effect of linear splicing on the circRNA assembly. The high complexity set also enables us to evaluate the effect of multiple overlapping circRNA s on circRNA assembly.* ***Reproduced from Stefanov et al. (2022)***

3.2.3 Polyestercirc

In order to provide an unbiased test, RNA-seq libraries containing linear and circular transcripts with multiple isoform overlap sequences must be simulated. The commonly used RNA-seq simulation tools are not designed with the intention of handling circRNA transcripts [121, 56, 122]. The simulation strategies previously used for the circRNA RNA-seq reads [115, 116] do not adequately represent real data. Some RNA-seq simulation tools were designed with the purpose of simulating circRNA reads [115], but the functionality is lacking in terms of realistic edge effect, sequencing errors, and GC bias. In this tool, ”sequence bias” is based on random selection of read start sites, which does not mimic real data. The simulation performed in [116] uses polyester by providing linearised circRNA transcripts extended by the length of a read(L) minus one nucleotide (L-1). However, this strategy does not consider circRNA-derived fragments that map to the BSJ and have a length longer than a read on either side of the junction.

Polyester is a tool that simulates RNA-seq reads following a negative binomial model of count distribution to produce replicates with a realistic differences [121]. I have made modifications to the original code of Polyester to create a new tool, named polyestercirc, which focuses solely on the simulation of circular RNA circRNA. This novel tool is designed to be used in conjunction with the original Polyester to generate a simulated library with reads corresponding to both linear and circular transcripts. The simulation parameters are optimised to generate the libraries described in Section 3.2.1.

The adjustments to the Polyester code are implemented with the assumption that circular RNAs (circRNAs) will experience at least one hydrolysis event during RNA fragmentation. The initial break point of the fragmentation is chosen at random along the sequence of the circle. To better align with the circular nature of circRNAs, I've disabled the edge effect model for simulating circular reads.

The recommended approach for using Polyester entails incorporating GC-bias by adapting the number of simulated transcripts based on their GC-content. To improve the realism of GC-bias modeling, I have modified the process to sampling transcripts from an initial set as Bernoulli trials with probability based on GC-content [52].

I simulate circular and linear transcripts separately and then merge them into one library. Linear transcript sequences are selected as described in Section 3.2.1. To mimic the noise caused by the presence of nascent RNA, I added the full sequence of the gene to the set of linear transcript reference. To simulate circRNA enrichment, I decrease the expression values of the linear transcripts by a factor of five and increase circular transcript levels by a factor of 4.5. The final library parameters that are selected to fit the requirements of all tools participating in the performance benchmark (see Section 3.5.1) are shown in Table 3.4.

	Rep1 control	Rep2 control	Rep1 treated	Rep2 treated
2x75	26,964,485	26,937,052	25,760,332	25,752,251
2x250	27,011,990	27,005,653	25,863,703	25,855,256

Table 3.4: **Information on simulated libraries type and depth.** *To accommodate the requirements of all tools, libraries were simulated in: 1) replicates; 2) pair-end; 3) two types of read length 4) 5 time decrease in the linear transcript abundance and 4.5 times increase in circular transcript abundance (based on empirical observations).* ***Reproduced from Stefanov et al. (2022)***

3.3 CYCLeR pipeline

3.3.1 Selection of a reliable BSJ set

The identification of novel BSJs hinges on the capability of an alignment tool to map chimeric reads. Different mapping strategies and heuristics can detect distinct sets of chimeric reads per sample. To ensure a reliable staring set of BSJ sites, I employ a combination of CIRI2 and CIRCexplorer2 as input. Their capability for detecting BSJ depends on the mapping tools incorporated into their pipeline, namely BWA-MEM and STAR for chimeric detection, respectively.

CYCLeR also allows for any other BSJ set to serve as input via a TSV file. A key input for the functionality of *CYCLeR* is a matching set of BSJs from control ribo-depleted and circRNA-enriched libraries. This combination allows for an easy filtering of false positive BSJ predictions. The option of merging different sources of BSJs input gives an added advantage to *CYCLeR*, as it has been shown that using a combination of BSJ identifications tools improves the quality of the prediction [13, 123] (see Table 3.5).

3.3.2 Creation of circular splice graphs

A necessary first step for the assembly of transcripts is to create a comprehensive splice graph. The process of constructing a splice graph is best explained with an example taken from preliminary simulated data. The steps of constructing a splice graph for the 5-HT2A *D.* melanogaster gene are clearly illustrated in the figure 3.4. The mapping of the reads is performed with STAR [34]. Then the reads assigned to the circRNAs loci are extracted with SAMtools [124] to eliminate unnecessary reads that slow the computation. Reads originating from linear and circular transcripts pass this filter. The exons and splice junctions are arranged in a splice graph using the SGSeq R package [120]. Processing of chimeric reads causes mapping artefacts and erroneous counting of the RNA-seq fragments. I resolve the mapping artifacts by correction using the BSJ information as well as known linear annotation. Next, I perform an exon read recount with the RSubread package [125]. DEXSeq R package [46] is used to eliminate the features depleted after circRNA-enrichment. I modified the standard workflow of DEXSeq to bypass the normalisation step. I opted for RPKM normalisation because it enables corrections to the effective length sequence, accounting for the drop in read coverage around BSJ sites. *CYCLeR* provides optional GC correction based on the GC content models used in the polyester package [121]. If replicates are unavailable for any of the two conditions, *CYCLeR* uses a direct comparison of the average values to eliminate linear-specific features. For more information on the type of the graph used, see Figure 3.5.

3.3.3 Reconstruction of circRNA transcripts

The next step in the *CYCLeR* workflow is a stepwise reconstruction of possible transcripts using a tailor-made graph algorithm. Graph algorithms are a common approach to processing of a splice graph [48]. In *CYCLeR*, I utilise a greedy algorithm for the iterative reconstruction of transcripts, aimed at minimizing the occurrence of false-positive assembled sequences. The starting point of the algorithm is the comprehensive splice graph, created in the preceding steps. The process begins by selecting the exon with the lowest abundance. Subsequently, *CYCLeR* proceeds to identify the maximum flow through this exon within the splice graph and

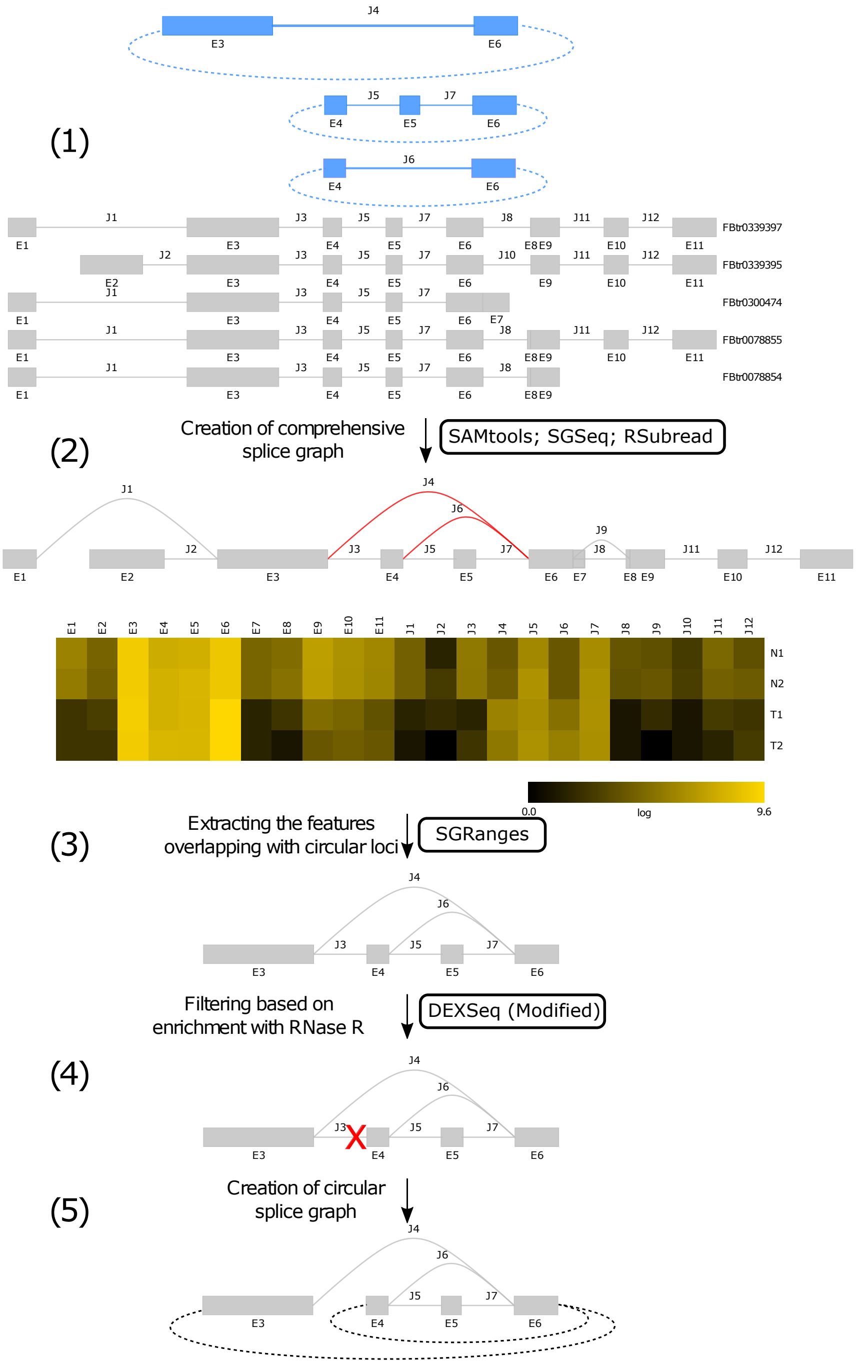

(1)
J4
E3
E6
J5
J7
E4
E5
J6
FBtr0339397
FBtr0339395
FBtr0300474
FBtr0078855
FBtr0078854
J1
J2
J3
J8
J9
J10
J11
J12
E1
E2
E7
E8
E9
E10
E11
(2)
Creation of comprehensive splice graph
SAMtools; SGSeq; RSubread
N1
N2
T1
T2
0.0
log
9.6
(3)
Extracting the features overlapping with circular loci
SGRanges
Filtering based on enrichment with RNase R
DEXSeq (Modified)
(4)
X
(5)
Creation of circular splice graph

Figure 3.4: **Construction of the circle-specific splice graph of the 5-HT2A gene.** *(1) Reference linear and circular transcripts are used for sample simulation, (2) comprehensive splice graph is created using SGSeq, the features for 4 samples are quantified and can be represented in a heatmap(N (non-treated), T (treated); the features that have no previous annotation are marked red), (3) a subgraph with only features located between the BSJ start/end boundaries is extracted, (4) filtering based on depletion of features in circRNA enriched samples (T) (5) Final splice graph creation.* ***Reproduced from Stefanov et al. (2022)***

outputs the corresponding sequence. After reconstructing the circular transcript, *CYCLeR* subtracts the abundances of the

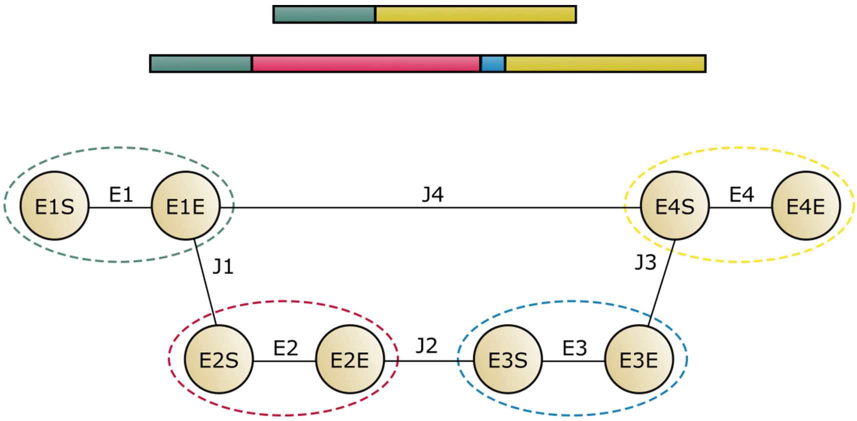


Figure 3.5: **Algorithmic representation of the splice graph** *In graph theory terms, the graph itself is a monopartite graph, in which for every exon or retained intron, there are two nodes corresponding to the 5'and 3' sites of the feature. Exon 3'-nodes can be connected to 5'- nodes via edges that represent splice junctions. The edges of the graph hold the information of the coverage of the features. This method of representation of splice graphs is somewhat similar to the approach used in StringTie. Information of the BSJ sites and coverage is handled separately.*

corresponding exons from their respective features in the original graph. Any features that become fully depleted as a result of this subtraction are eliminated. This operation is repeated until no further transcripts can be reconstructed. For an illustrative example, see Figure 3.6. [52].

3.3.4 CircRNA transcript quantification

For drawing conclusions based on transcript abundance, downstream analyses typically involve a single value for read counts per transcript. The preferred approach for transcript quantification in recent years have been EM-based tools [56, 54]; for more information, see Section 2.4.2. *CYCLeR* leverages the strategy used in sailfish-cir and enhances it by incorporating a functional circular RNA (circRNA) assembly step. For every assembled circular transcript, *CYCLeR* produces a pseudo-linear reference by adding a string matching the sequence at the end of an edge exon. For simultaneous quantification, the newly created pseudo-linear reference needs to be combined with the linear transcript reference. The final step of abundance estimation is performed with kallisto [54], using the newly generated comprehensive reference. The pseudo-linear adjustment consists of the addition of the size of a k-mer to be used in the quantification step, as illustrated in Figure 3.7. This makes kallisto

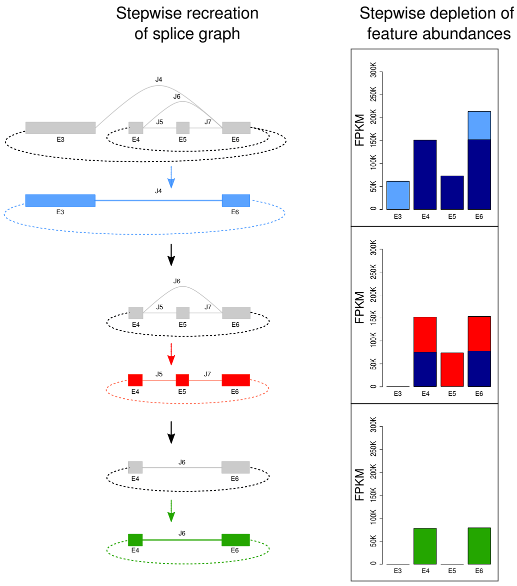

Stepwise recreation of splice graph
Stepwise depletion of feature abundances
J4
J6
J5
J7
E3
E4
E5
E6
FPKM
0
50K
100K
150K
200K
250K
300K

Figure 3.6:* Circle transcript reconstruction within *CYCLeR* for the example of the 5-HT2A gene.** *Starting with the full splice graph for the entire gene locus and its respective exon abundances (see top line, left, in grey for the graph and the FPKM-plot at the top right),* CYCLeR *extracts the circle-specific sub-graph corresponding to splice-site junction J4 which falls between a BSJ-start and a corresponding BSJ-end coordinate (see second line from top left, in blue for the graph). This blue sub-graph corresponds to a single circular splicing isoform. This blue sub-graph and the corresponding exon abundances are subsequently subtracted from the original, full splice graph (see graph at the top left and middle FPKM-plot) to yield the remaining splice graph (third line from top, in grey). Similarly to before,* CYCLeR *then extracts the next circle-specific sub-graph, this time corresponding to a* back-splice junction*-spanning splice-site junction between exon E6 and E4 (fourth line from top, in red). This sub-graph provides evidence for a circular splicing isoform comprising three exons E4, E5 and E6 (note the different exon abundances). The quantities corresponding to this circular isoform are subsequently deleted from the remaining grey splice graph, resulting in a subgraph (second line from bottom, left, grey) that corresponds to another circle-specific graph, this time comprising only exons E4 and E6, but not E5 (bottom line, in green). This sub-graph and its abundances provide evidence for a single circular isoform (bottom FPKM-plot).* ***Reproduced from Stefanov et al. (2022)

aware of the additional mapping possibilities around the back-splice junction (BSJ). kallisto also requires an adjustment to increase the effective length of the sequence, achieved by the addition of pseudo-random nucleotides (not similar to the reference sequences).

3.4 Code and Documentation

CYCLeR is available at https://github.com/stiv1n/CYCLeR. The repository contains information about a trial run of the core script, as well as all needed command line tools, both provided via Docker.

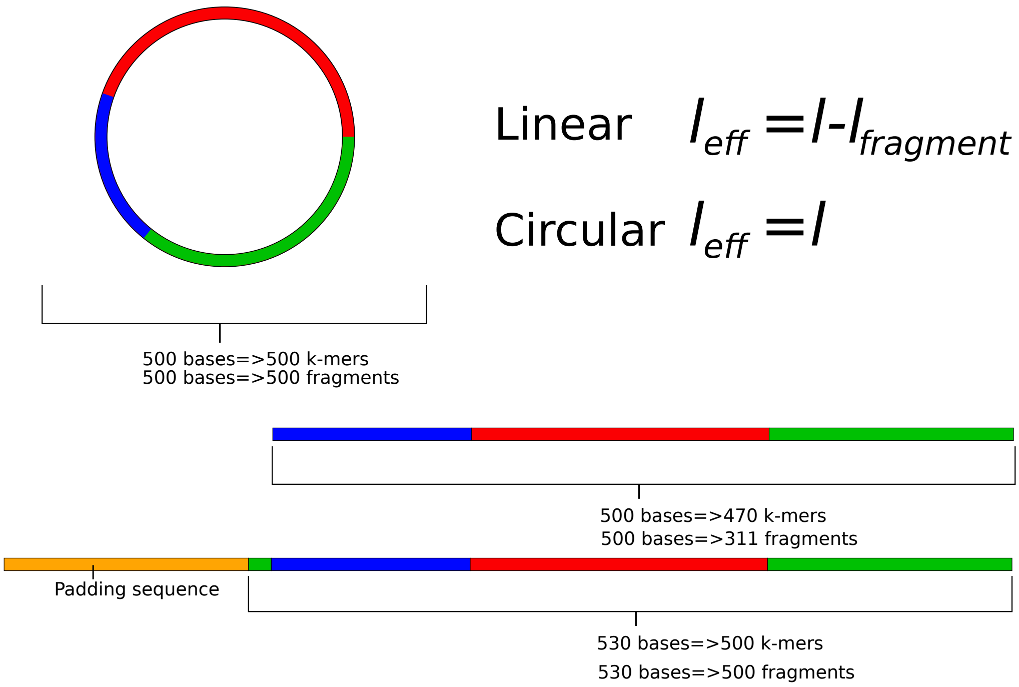


Figure 3.7: **Creation of a pseudo-linear reference sequence.** *Comparison between an example circular and a linear transcript of the same length (l)=500 bp. Considering a k-mer size of 31 bases, the circRNA produces 30 more unique k-mers compared with the linear transcript of the same size. Therefore, an extra of 30 bases need to be added to the pseudo-linear transcript when an isoform quantification is used that is based on pseudo-alignments. Additional padding sequence of pseudo-random nucleotides is added for adjustment of effective length(l_{eff}) equal to one insert(*aka *fragment) length(l_{fragment}).* ***Reproduced from Stefanov et al. (2022)***

This section is a copy of the GitHub page of *CYCLeR*.

3.4.1 Working with CYCLeR

CYCLeR is a pipeline for reconstruction of circRNA transcripts from RNA-seq data and their subsequent quantification. The algorithm relies on comparison between control total RNA-seq samples and circRNA enriched samples to identify circRNA specific features. Then the selected circRNA features are used to infer the transcripts through a graph-based algorithm. Once the predicted transcript set is assembled, the transcript abundances are estimated through an EM algorithm with **kallisto**. **CYCLeR** takes as an input BAM files and back-splice junction (BSJ) files and outputs transcript infomation in different formats, including a FASTA output for abundance estimation.

3.4.1.1 Installation of CYCLeR

3.4.1.1.1 Command line tools needed

The computation steps prior and post **CYCLeR** run are most efficiently run on HPC. It is very likely that any HPC in biological institute already has most of those tools installed. Just in case, a **Docker** image containing all the tools is provided. NOTE: prior to running **Docker** image, make sure that ***Docker** is indeed installed and working: https://docs.docker.com/get-started/

- **STAR** - https://github.com/alexdobin/STAR
- **samtools** - https://sourceforge.net/projects/samtools/files/samtools/
- **kallisto** - http://pachterlab.github.io/kallisto/download
- **bwa** (needed for CIRI2) - http://bio-bwa.sourceforge.net/bwa.shtml
- **CIRI2** - https://sourceforge.net/projects/ciri/files/CIRI2/
- **CIRCexplorer2** - https://circexplorer2.readthedocs.io/en/latest/

3.4.1.1.2 Docker image with all command line tools

```
sudo docker pull stiv1n/cycler.prerequisites
```

3.4.1.1.3 R test run installation

For a test run, I suggest using a **Docker** container. There, all test input and all dependencies are provided. The **Docker** use requires you to mount a volume - a working directory (**local_dir**) where the output and input would be stored. This container uses **RStudio server** and required login. In this case, the username is *rstudio* the password is *guest*. Open a browser and type `localhost:8787`.

```
sudo docker pull stiv1n/cycler
sudo docker run --rm -ti -e PASSWORD=guest \
-v <local_dir>:/usr/workdir -p 8787:8787 stiv1n/cycler
```

3.4.1.2 Full documentation

For more information on installation, pre-processing and core tool run please see the vignette[1] and the manual[2].

[1] https://raw.githubusercontent.com/stiv1n/CYCLeR/main/CYCLeR_workflow.pdf

[2] https://raw.githubusercontent.com/stiv1n/CYCLeR/main/CYCLeR.pdf

3.4.1.3 CYCLeR

To test the tool, use the R script[3] provided and just run it in the RStudio server.

3.4.1.4 Quantification

The final step of the **CYCLeR** pipeline is running **kallisto** with the newly assembled transcriptome.

```
sudo docker run  \
  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  kallisto index -i kall_index -k 31 for_kallisto.fa
sudo docker run  \
  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  kallisto quant -i kall_index \
    -o ./ <sample_name>_1.fq <sample_name>_2.fq
```

3.5 Benchmark design

At the time of writing the thesis, there is no tool that has the exact same functionality as *CYCLeR*. The differences between the main competitor tools are described in detail in Table 3.5. Therefore, I needed to benchmark the assembly and quantification of *CYCLeR* separately. The first step of the *CYCLeR* pipeline is the assembly of the full-length circular isoforms. The tools that perform such task are grouped into class IV. In principle, CIRCexplorer 2 from class II can also be included in this benchmark, since it produces potential transcripts derived from all possible combinations of predicted AS events [52].
The second part of the benchmark that focuses on circRNA abundance estimation is more challenging. The *CYCLeR* pipeline performs the simultaneous quantification of linear and circular transcripts. In contrast, there are multiple alternative approaches taken when designing the tools described in Table 3.5. Almost all of the tools in the benchmark focus only on the quantification of circRNAs. Among the tools in the benchmark, the only tool except *CYCLeR*, that performs simultaneous quantification of linear and circular transcripts is sailfish-cir. sailfish-cir, however, does not

[3] https://raw.githubusercontent.com/stiv1n/CYCLeR/main/docker_test.R

perform novel isoform assembly. The rest of the quantification strategies, described in Table 3.5, calculate the abundance based on the reads mapped to the BSJs. From those tools, only CIRI-full provides complete isoform abundance information. This forced me to perform multiple separate quantification benchmarks, where the primary focus is the advantage of the general strategy that *CYCLeR* uses, as opposed to highlighting performance statistics with a single unified benchmark. The parameters used to run the tools are summarised in Figure 7.6 and correspond to the suggested parameters from the corresponding manuals. Information on reference genomes and annotations can be found in Table 7.1.

3.5.1 Benchmark with simulated data

It is common practice to benchmark tools for transcriptome assembly and quantification via a set of dedicated simulated data. The optimal input for *CYCLeR* is an RNA-seq dataset that combines ribo-depleted RNA and circRNA-enriched libraries. The circRNA feature detection module performs optimally with replicates. Such dataset also fits the requirements of CIRCexplorer 2. Class IV methods benefit from long library inserts with 250 bases sequenced on both ends. As discussed in Section 2.3.2.2, rolling-circle amplification is an issue for RNA-seq libraries. It should be noted that the CIRI-full algorithm is designed to handle rolling circle cDNA product, but the tool can also perform well with a library generated through RNA fragmentation. To avoid unpredictable issues related to rolling-circle amplification, I decided to focus the benchmark solely on library preparation involving RNA fragmentation. To best fit all requirements of the tool, I simulated two types of RNA-seq library: one library with a median fragment length of a 280 bp and 75 bp sequencing and one library with a median fragment length of 500 bp and 250 bp sequencing. From both library fragment setups, I have simulated the full set of required libraries; details in Section 3.2.1, as well as Tables 3.2, 3.3 and 3.4 and Figures 3.2 and 3.3. I employed *CYCLeR*, CIRCexplorer2 and CIRI-full on both types of simulated data sets with parameters set suggested in their respective user manuals. Note that I do not include sailfish-cir in this benchmarking due to its inability to identify new transcripts. Downstream steps of an analysis, such as

Software	CircRNA feature selection	*De novo* feature identification	Transcript reconstruction	Transcript quantification	Flexibility*
sailfish-cir	exons within circRNA boundaries selected based on known linear annotation	—	available linear annotation is used to infer AS	EM quantification based created pseudo-linear reference	yes
CIRIquant	—	FSJ within circRNA boundaries selected based on HISAT [126] mapping and StringTie [48] assembly of circRNA enriched libraries	—	fitting circRNA levels to a gaussian mixture model, combining circRNA enriched and total RNA-seq data	yes
CircAST	—	exon boundaries detected based on splice junctions derived from Tophat2 [33] mapping	minimum set of paths between BSJs that include all splice junctions	EM algorithm based on adjusted fragment length distribution	yes
CIRC-explorer suite	comparison between total and circle enriched RNA-seq libraries	RABT assembly with Cufflinks [47] (StringTie [48] in latter versions), based on TopHat2 (HISAT [126]) mapping	statistical test to determine AS events and reconstruct all potential isoforms	CLEAR [118] add-on quantification of circRNA as a ratio based on the levels of the most predominant equivalent linear transcript	no
CIRI suite	exon selection based of pair-end reads	internal Perl script detecting junctions and retained introns from BWA [35] mapping	AS events are detected with statistical test based on difference of coverage between exons	transcript quantification though iterative optimisation of exon abundances within a pre-constructed splice graph	no
CYCLeR	comparison between total and circle enriched RNA-seq libraries with DEXSeq [46] package	feature detection through SGSeq [120] package based on STAR [34] mapping	transcript reconstruction using a greedy algorithm on splice graph	EM quantification based created pseudo-linear reference	yes

Table 3.5:* Overview of relevant circRNA transcript reconstruction and quantification tools.** *Abbreviations used: AS - alternative splicing, BSJ* -back-splice junction, *EM - expectation maximisation; * indicates that the tool is fully compatible with various BSJ identification tools.* ***Reproduced from Stefanov et al. (2022)

quantification, rely on prior knowledge of the isoform sequences. Consequently, both sensitivity and precision of the assembly are equally important for a robust analytical scheme. To emphasise that *CYCLeR* provides a good balance between those two measures, I also calculate the F-score - the harmonic mean of the sensitivity and precision. Sensitivity, precision, and F score are calculated in the usual manner [52]:

$$Sensitivity = \frac{\#\ of\ correctly\ predicted\ transcripts}{\#\ of\ all\ simulated\ transcripts}$$

$$Precision = \frac{\#\ of\ correctly\ predicted\ transcripts}{\#\ of\ all\ predicted\ transcripts}$$

$$F\text{-}score = 2 \cdot \frac{Sensitivity \cdot Precision}{Sensitivity + Precision}$$

Among the tools involved in the benchmark, only CIRI-full has an algorithm for quantification of different circRNA isoforms. The benchmark criterion is calculated as Pearson correlation product based on the estimated values of correctly assembled transcripts. Note that the calculations differ with regard to the difference in the nature of the output.

For *CYCLeR* calculation is done as follows:

$$corr(Assigned\ reads\ per\ transcript,\\ Simulated\ reads\ per\ transcript)$$

and CIRI-full as:

$$corr(Assigned\ BSJ\ reads\ per\ transcript,\\ Simulated\ number\ of\ transcript\ copies)$$

3.5.2 Benchmark with real data

3.5.2.1 Benchmark with Nanopore data

As stated in Section 2.3.4.2, full-length circRNA isoforms can be detected by using long-read sequencing. However, such approaches come with known and unknown biases. Furthermore, circRNAs enrichment for such protocols is far superior to what is common for RNA-seq libraries. Therefore, results from such protocols cannot serve as a golden standard for a benchmark of tools working with RNA-seq. Nonetheless, isoforms recovered from long-read studies can serve as a partial verification of the assembly of short-read data.

Considering this, the set of experiments used in the CIRI-long [43] publication provides suitable data for a benchmark supplying matching RNA-seq and Nanopore experiments, fitting the requirements of all tools. I compared the outputs of *CYCLeR* CIRI-full and CIRCexplorer2 in the context of the results from CIRI-long.

Various tools come with distinct default thresholds and parameters for assembly, and these may not be well-suited for the high sequencing depth of this study. In order to standardize the assembly parameters across the tools, I applied a threshold that considers only circRNAs with a minimum of five back-splice junction (BSJ)-spanning reads. [52].

3.5.2.2 Benchmark with *D.* melanogaster data

I wanted to showcase the advantage of transcript assembly and EM abundance estimation for exploratory analysis of real data. However, *CYCLeR*, among other tools, requires circRNA-enriched library to function well. Yet, there has been no directed effort to create a large database of circRNA-enriched libraries. Nevertheless, this gave me an opportunity to test the limits of *CYCLeR* in a study where circRNA-enriched libraries are available for only a few key time points. I selected *D.* melanogaster as a model organism for this benchmark due to the availability of RNase R treated samples from S2 cell line (derived from late stage embryo), early embryo and mature fly head (GSE69212, GSE55872) [9, 127]. I used *CYCLeR* to assemble circRNA transcripts from those data points and merged all, together with linear annotation, into a unified reference. Previously, the Lai lab performed an exploratory study on 103 *D.* melanogaster samples using a tool from class I [37]. I performed an analogous study comparing the results of *CYCLeR* to the results of representative methods of class I and class III.

The BSJ identification module of CIRCexplorer 2 has been shown to outperform the majority of the tools in class I [119], hence, I selected it as a representative. The only representative of class III that can perform well without a dedicated circRNA enriched library is sailfish-cir, leaving it the only choice for this comparative study. As class IV methods require circRNA-enriched libraries for optimal performance, they are not part of this study.

I normalise BSJ counts from CIRCexplorer 2 as counts-per-billion (CPB) and convert abundance of *CYCLeR* and sailfish-cir to RPKM. I perform variance-stabilisation with the DESeq2 package, and use the resulting values to calculate Spearman's rank correlation calculation to produce an adjacency (similarity) matrix. I used the adjacency matrix as an input for topological overlap matrix calculation based on the WGCNA package [64], without transforming it with an adjacency function.

3.5.2.3 Benchmark with qPCR

qPCR with primers converging on the BSJ is a useful tool for a rough estimate of the amount of certain circRNA species; specifically for mono- and diexonic circRNAs. The downside of such measurement is that it does not give any information about the alternative splicing of the transcripts sharing a BSJ. Furthermore, qPCR is arguably more accurate quantity measurement than abundance estimation based on levels of the transcript in an RNA-seq library, due to the fact that the full-length transcript is heavily influenced by sequencing biases; mainly GC-bias. As a consequence, benchmark with qPCR theoretically benefits tools of class I and class IIIa which focus on the BSJ read coverage and lack the ability to detect alternative splicing isoform mapping to the same BSJ. The tools from class IIIa perform adjustments based on the levels of linear transcripts, using the FSJs that circular and linear transcripts share as a reference point. This strategy is the most robust when a circRNA has just one FSJ. Thus, the values from publicly available benchmarks focus only on circRNAs with two exons.

I performed a comparative study using the data from [118]. That includes PA1 cell line libraries from the datasets GSE75733 and GSE73325 [50, 128, 118] and use the qPCR values reported in [118]. I omitted the circRNA cases with reported AS, because there is no practical way to match the values of *CYCLeR* to the qPCR values. The values are calculated as Pearson correlation product based on qPCR and *CYCLeR* abundance estimate averaged between two PA1 replicates. For comparison, the values reported in [118] are used. CIRIquant is not included in this comparison, due to its inability to analyse single-end RNA-seq data.

3.6 Results

3.6.1 CircRNA transcript assembly from simulated data

The tool CircAST consistently failed to conclude computations due to virtual memory issues, despite being provided with up to 400 GB RAM. This persistent issue compelled me to exclude CircAST from the benchmark.As mentioned earlier, CIRI-full requires a circRNA-enriched library for optimal use; therefore, only the results from such a library are shown in the benchmark. I merged the data from two replicates into a single output for each tool. As explained in Section 3.2.2, the results are compared to two reference sets with different complexity, referred to as reference set and high-complexity reference set. We can see the sensitivity and precision and F-score plot based on the reference set in Figure 3.8 and the corresponding results of the high complexity set in Figure 3.9. For both reference sets and for both RNA-seq read lengths, *CYCLeR* clearly outperforms over the competition and provides good balance across a variety of accuracy evaluations.

3.6.1.1 CIRI-full results

The most straightforward results to interpret are those from CIRI-full. In Figure 3.8, we can observe that CIRI-full achieves relatively high precision, but its sensitivity is limited. Such an outcome is entirely predictable, considering that by design, the algorithm outputs the full sequence of circRNAs only in cases when putative exons are covered by reads spanning the BSJ. This strategy imposes a limitation on the algorithm based on the insert size of the library.

For the 2x75 bp dataset, CIRI-full recovers mainly the sequences of loci with a single circRNA isoform. Given that such loci have very low complexity, CIRI-full has deceptively high precision for the 2x75 bp dataset. This deceptiveness translates to the high complexity benchmark, see Figure 3.9, where CIRI-full manages to maintain an acceptable precision value for the 2x75 bp dataset.

The 2x250 bp datasets show the actual capabilities of CIRI-full. The sensitivity increases as the library read length and insert size are optimized for the algorithm. However, the precision drops due to the increased complexity of the alternative splicing landscape typical for longer circRNAs. As a conclusion from the results described above, CIRI-full is incapable of handling long circRNAs.

3.6.1.2 CIRCexplorer2 results

The primary purpose of CIRCexplorer2 is the detection AS of circRNAs and providing potential transcript sequences as an additional output as a combination of the detected AS events per BSJ. While this approach leads to high sensitivity, the precision is low, as a result of unrestricted AS event combinations. Since the feature selection of CIRCexplorer2 is quite straightforward, it also benefits from an increase in read length similar to CIRI-full. The permissive filtering of transcriptomic features allows CIRCexplorer2 to outperform the competing tools in sensitivity in the high complexity benchmark. Nevertheless, the precision and correspondingly the F-score in both high and low complexity benchmarks indicate that CIRCexplorer2 is an unreliable tool to use for circRNA assembly.

3.6.1.3 *CYCLeR* results

The F-score in Figure 3.8 shows an overwhelming advantage of *CYCLeR* to CIRI-full and CIRCexplorer2. The sensitivity measures for *CYCLeR* between the 2x250bp and the 2x75bp datasets can be considered unusual. Tools developed for transcriptome analysis typically exhibit improved performance with longer read lengths. However, the normalisation and the specific the effective length adjustment of the feature quantification in *CYCLeR* is made in a way that benefits reads that span a single FSJ. Since 250 bp reads often map to multiple FSJ, the effective length adjustment is needlessly overcompensating.

CYCLeR excels in the sensitivity measure for the 2x75 libraries, while CIRCexplorer2 holds a slight advantage for the 2x250 libraries. However, the substantial superiority of *CYCLeR* based on precision unquestionably establishes it as the overall superior tool.

3.6.1.4 Comments

These results shown in Figures 3.8 and 3.9 can be explained by two key features that give *CYCLeR* an advantage—transcript feature selection and the graph algorithm. The biggest advantage of

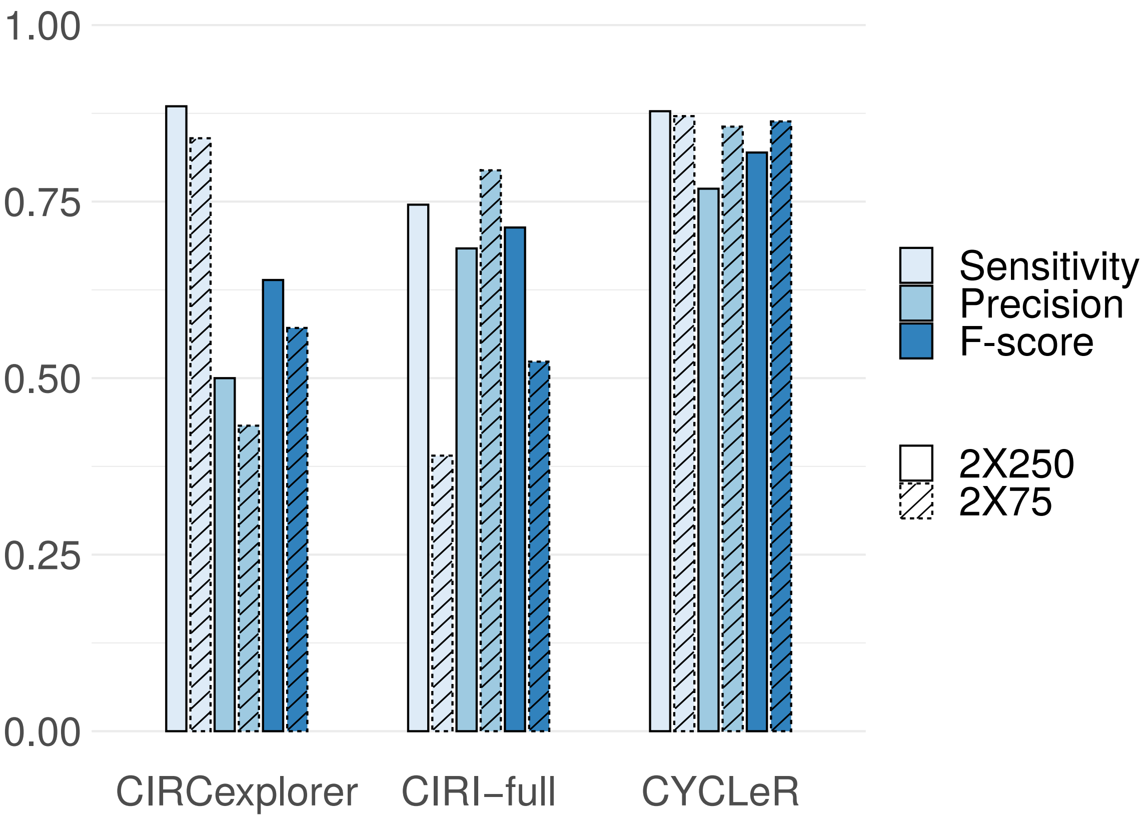


Figure 3.8: **Comparative benchmarking of *CYCLeR* and comparable tools.** *Bar plot of sensitivity precision and F-score of* CYCLeR *and different existing tools based on the simulated reference dataset. The superior F-score for* CYCLeR *shows a good balance between sensitivity and precision.* CYCLeR *outperforms CIRI-full on all metrics. CIRCexplorer2 matches the sensitivity of* CYCLeR*, but the number of false positive assemblies shown by the precision measure makes CIRCexplorer2 an unreliable choice. It is important to note that* CYCLeR *output is only marginally affected by the library read length.* ***Reproduced from Stefanov et al. (2022)***

CYCLeR can be attributed to the robust selection of transcript features. The key feature of this is the use of replicate data and the statistical model of DEXSeq to filter out exons and junctions that do not participate in circRNAs. That way, *CYCLeR* avoids the obstruction caused by residual linear RNAs. As a consequence, *CYCLeR* can consistently outperform the competition in cases of low coverage loci and loci with unannotated circRNA transcript features. CIRCexplorer2 (using Cufflinks/ StringTie) and CIRI-full (using customized scripts) do not have as secure a way to filter features belonging to linear RNA, forcing them to either rely heavily on annotation or focus only on features with high read coverage. An example of this can be seen in Figure 3.10. The advantage of *CYCLeR* in feature selection can also be attributed to the increased starting set of BSJs, since *CYCLeR* uses a combination of class I

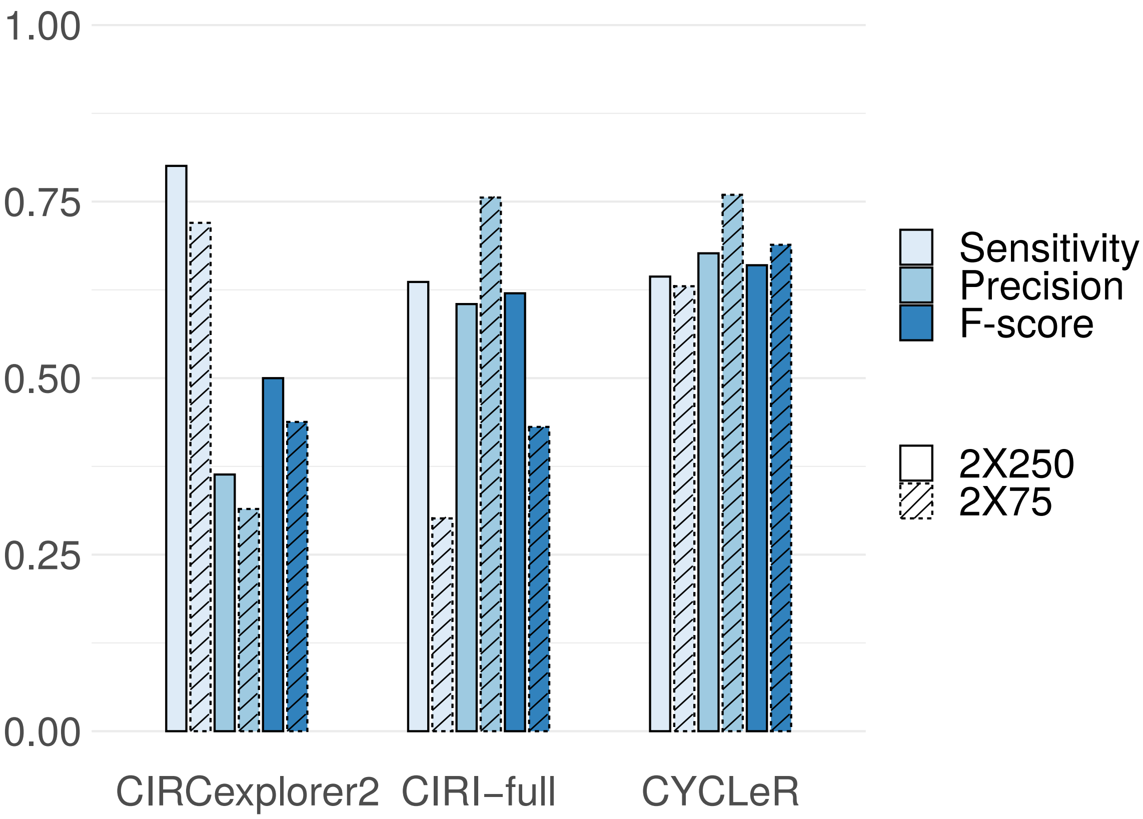


Figure 3.9: **Benchmark with the *High complexity* dataset.** *This barplot depicts the* sensitivity, precision and F-score *of the assembled transcripts by* CYCLeR *and comparable tools in the benchmark for reference set of simulated data. The advantages of* CYCLeR *compared to other tools are apparent. The superior F-score of* CYCLeR *shows a good balance between sensitivity and precision.* CYCLeR *outperforms CIRI-full on all metrics. CIRCexplorer2 has an output with higher sensitivity than* CYCLeR, *but the number of false positive assemblies shown by the precision measure makes CIRCexplorer2 an unreliable choice. Note that* CYCLeR *is only minimally affected by the library difference in library read length.* ***Reproduced from Stefanov et al. (2022)***

tools as input. Another source of high precision for *CYCLeR* is the graph algorithm that is custom-made for the assembly of the circRNA transcript. The parameters hard-coded in the algorithm ensure a secure step-wise reconstruction of the transcripts, benefiting precision over sensitivity. This explains the overall high F-score across all measurements.

3.6.2 CircRNA transcript quantification from simulated data

Quantification of circRNA isoforms is a problem mainly in the cases of multiple overlapping isoforms. Therefore, the quantification benchmark focuses only on the high-complexity data set. As explained earlier, only class IV tools can participate in this

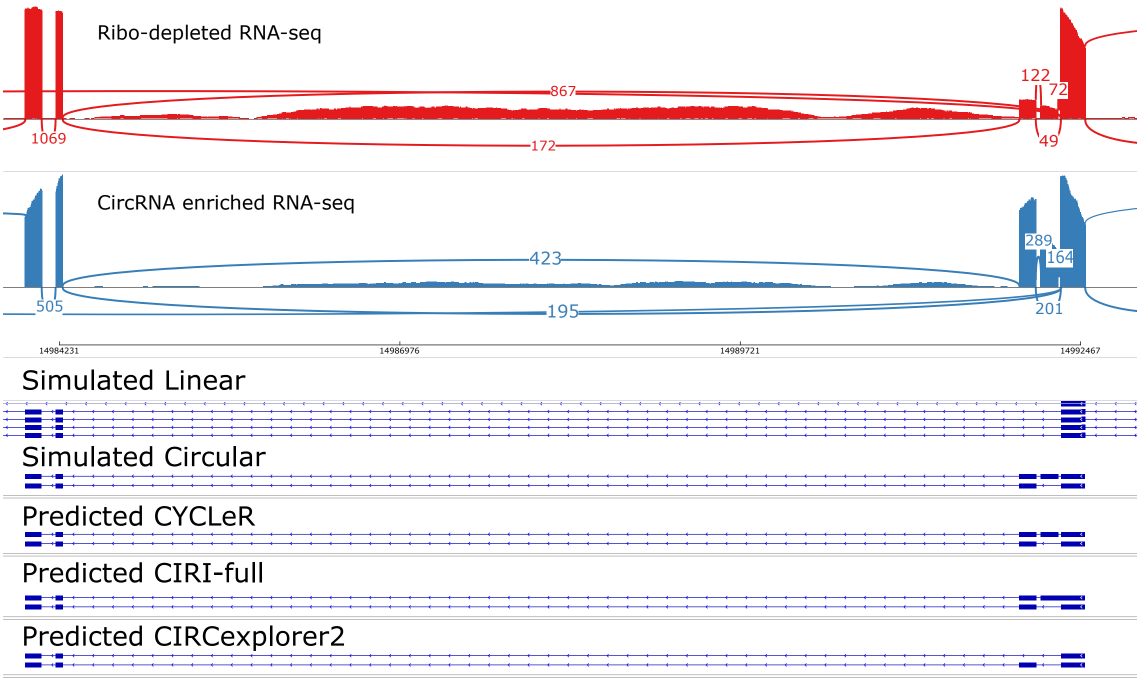


Figure 3.10: **Comparison of the assembly of BSJ locus chr2L:14,983,950-14,992,506 in *D.* melanogaster.** *The Sashimi plots show the STAR mapping of simualted data for the exons and FSJ encompassed by the BSJ sites of the circle in chr2L:14,983,950-14,992,506. The plot shows comparison between total ribo-depleted RNA-seq simulation and circRNA enriched RNA-seq. The sequence mode used is 2 x 250 (leading to less observable biases that in Figure S1). This example was selected to show the advantage of* CYCLeR *in handling unannotated exons. While all tools can handle identification of one of the unannotated exons, the second exon is accounted for properly only by* CYCLeR. *We observe that the CIRCexplorer2 assembly is biased by the provided linear annotation. The CIRI-full assembly disregards one of the FSJs and assembles an erroneous full exon.* CYCLeR *not only correctly identifies all exons and splice junctions, bus also manages to properly manage the AS event and reconstructs the correct isoforms.* ***Reproduced from Stefanov et al. (2022)***

benchmark, as all other tools quantify circRNA with a disregard for alternative isoforms.

In Table 3.6 we can see the correlation of estimated values for circRNA transcripts levels versus simulated quantities in both ribo-depleted and circRNA-enriched data.

It is important to reiterate that *CYCLeR* is the only existing tool that simultaneously quantifies both known linear and newly assembled circular transcripts. Note that CIRI-vis [129] (the quantification module of the CIRI-suite) is not affected by the presence of linear transcripts in the same way as *CYCLeR*. CIRI-vis provides levels of the alternatively spliced circRNAs as a ratio of the BSJ-spanning reads assigned to each isoform. Therefore, linear RNA affects the results from CIRI-vis only at the assembly level.

For the 2X75 data the correlation statistics of CIRI-vis and *CYCLeR* are equivalent. However, we need to account for the low sensitivity of CIRI-full and elaborate that complex cases are therefore never considered by this tool. Therefore, *CYCLeR* can be regarded as an outperformer. The advantage of *CYCLeR* is more noticeable for the 2X250 data. While both tools show improved performance, *CYCLeR* is significantly superior. The difference should not be attributed solely to the quantification strategy, but also to the differences in reference provided by the assembly step. Specifically, correct quantification is hindered by false positive transcripts.

	Ribo-depleted		CircRNA enriched		
Tool	Replicate1	Replicate2	Replicate1	Replicate2	Type
CYCLeR	0.57	0.64	0.66	0.66	2 x 75
CIRI-vis	0.54	0.64	0.66	0.67	2 x 75
CYCLeR	0.84	0.85	0.87	0.88	2 x 250
CIRI-vis	0.67	0.66	0.76	0.74	2 x 250

Table 3.6: **Correlation of predicted versus simulated circRNA transcript counts.** *Correlation of predicted transcript abundances versus simulated. Correlations are based only on the values of correctly identified transcripts by both tools. The values are based on correlations for the transcripts of the high complexity set.* ***Reproduced from Stefanov et al. (2022)***

3.6.3 Comparative study with Long-read data

The CIRI-long protocol is capable of detecting circRNA transcripts beyond the capabilities of standard circRNA enrichment strategies used in the preparation of RNA-seq [43]. This makes CIRI-long a poor benchmark for false negative transcript reconstructions, due to its high sensitivity for low-abundance circRNAs. Thus, CIRI-long data cannot be used as a reference for a benchmark in the same capacity as a simulated data set. There are also some known biases of the CIRI-long strategy. It is important to note that ∼50% of the BSJs detected in the Illumina data are not detected in the CIRI-long data [43]. My assumption is that the 1000 nt fragment size selection limit of the CIRI-long protocol leads to a loss of long-length circRNAs. This also makes CIRI-long results a bad estimate for true positive transcript reconstructions.

The most meaningful statistical analysis that can be conducted is the assessment of prediction overlaps between the Illumina-based tools (*CYCLeR*, CIRI-full, CIRCexplorer2) and CIRI-long. To enhance the presentability of the statistics, the following terms are introduced:[52]:

1. **Verified Isoforms:** Isoforms confirmed by any Illumina-based tool and CIRI-long data.
2. **Shared Isoforms** *Verified* isoforms predicted by multiple Illumina-based tools.
3. **Unique Isoforms:** *Verified* isoforms predicted by a single Illumina-based tool.
4. **Unverified Isoforms:** Isoforms predicted by Illumina-based tools but lacking support from CIRI-long data.

CIRI-full is constrained by the maximum recoverable transcript length, a limitation coincidentally similar to CIRI-long. To ensure a fair and unified benchmark, I adjust all results to consider CIRI-full length limitation. The modified results are presented in Figure 3.11, with complete results available in Figure 3.13.

In Figures 3.11 A and B, we can see the comparison of the adjusted results. It is important to note that all three RNA-seq-based tools identify a set of unique transcripts, verified by CIRI-long. A key observation is that the output of CIRCexplorer2 contains an inordinate number of isoforms compared to *CYCLeR*or CIRI-full. While CIRCexplorer2 provides the highest number of verified isoforms, the number of unverified isoforms from CIRCexplorer2 is disproportionately high. *CYCLeR* has both higher number of verified and unverified transcripts than CIRI-full.

To fully understand the results, we must take into account the number of BSJs that a tool uses for the assembly shown in Figure 3.12. The exceptionally high number of transcripts reconstructed by CIRCexplorer2 cannot be explained by the input of BSJ alone. In fact, CIRCexplorer2 has the lowest number of unique BSJs. Even when we take into account the expected biases of the Nanopore approach, the number of transcripts in the output of CIRCexplorer2 is exorbitant. These results are in perfect agreement with the benchmark on simulated data and support the notion that CIRCexplorer2 produces a high number of false positive transcripts.

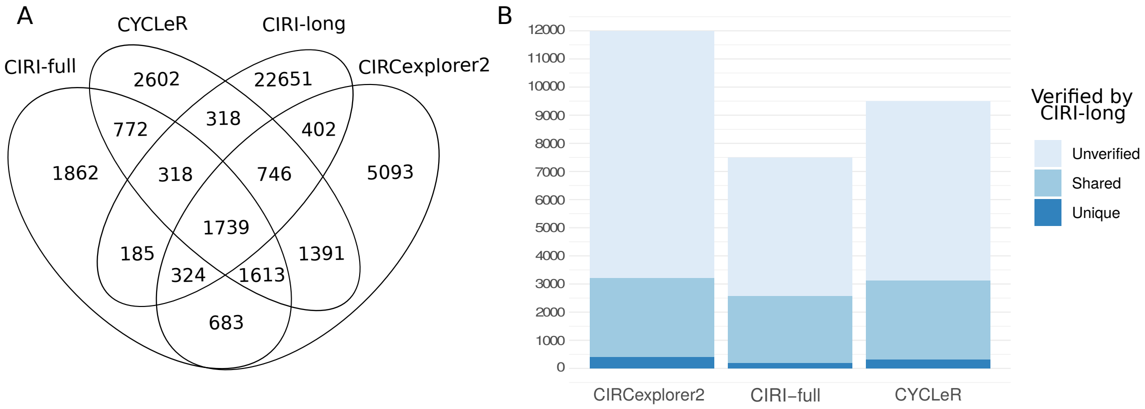


Figure 3.11:* Comparative study with CIRI-long data.** *(A) and (B) show the results of the comparison between Illumina-based methods and a Nanopore-based method. (A) shows a Venn diagram of the length adjusted (<2000) set of assembled transcripts for each tool. (B) is a bar graph representation of the same data, but with emphasis on overlapping regions from the Venn diagram. In figure B the assembled transcripts for each Illumina-based tool are divided into* verified *(by CIRI-long) or* unverified*. The latter are further subdivided into* unique *- the transcripts that are shared only by one Illumina-based tool and CIRI-long, and* shared *- the transcripts that are shared by two or more Illumina-based tools and CIRI-long. CIRI-full has the lowest transcript count in every category. This is due to the length limit of its underlying assembly based on the library insert size. When comparing CIRCexplorer2 and* CYCLeR*, we notice that CIRCexplorer2 has only ∼100 more* verified *transcripts, while simultaneously having ∼3000 more* unverified *transcripts. Based on the information provided by the simulated benchmark, it is a safe assumption that the extra isoforms produces by CIRCexplorer2 are mainly erroneous assemblies.* ***Reproduced from Stefanov et al. (2022)

Figure 3.11 also shows that *CYCLeR* has more verified transcript predictions than CIRI-full. Additionally, both tools have a ratio of ∼60% unverified isoforms, indicating comparable precision of *CYCLeR* and CIRI-full . It is also important to note that the number of unique BSJs for CIRI-full and *CYCLeR* is similar, indicating that the differences in the assembly are attributed to the algorithm and not the starting BSJ set. The combination of these results points to an advantage of *CYCLeR*over CIRI-full, most likely a consequence of the ability of *CYCLeR*to handle longer transcripts.

The superiority of *CYCLeR* over CIRCexplorer2 and CIRI-full is shown in the full results (without length adjustment), shown in Figure 3.13. An example of the advantage of *CYCLeR* can be seen in Figure 3.14, from which it follows that CIRI-full cannot match the recall of *CYCLeR* and CIRCexplorer2, due to the limitation of the length of the transcript. Furthermore, CIRCexplorer2 has a problem with false positive assembly of transcripts, leading to low precision. This makes *CYCLeR* the only remaining good choice when looking for a tool with

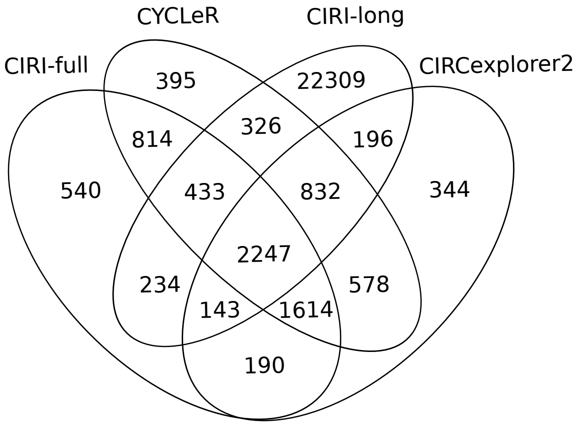


Figure 3.12: **Venn diagram of unique BSJ per tool in the benchmark versus CIRI-long Nanopore data.** *The assembly of each tool is dependent of the set of input BSJs. This plot complements the assembly of the transcript (shown in Figure S17 and Figure 6 in the manuscript)and sheds light on the differences in prediction. The CIRI-full and CIRCexplorer2 BSJ that are not part of the* CYCLeR *output derive from loci that failed the BSJ enrichment requirement. The BSJs that are unique for* CYCLeR *are BSJs identified by CIRI2, which belong to circRNAs longer than the CIRI-full detection limit. CIRCexplorer2 has unique set of BSJs disproportional to the number of unique transcripts assembled.* ***Reproduced from Stefanov et al. (2022)***

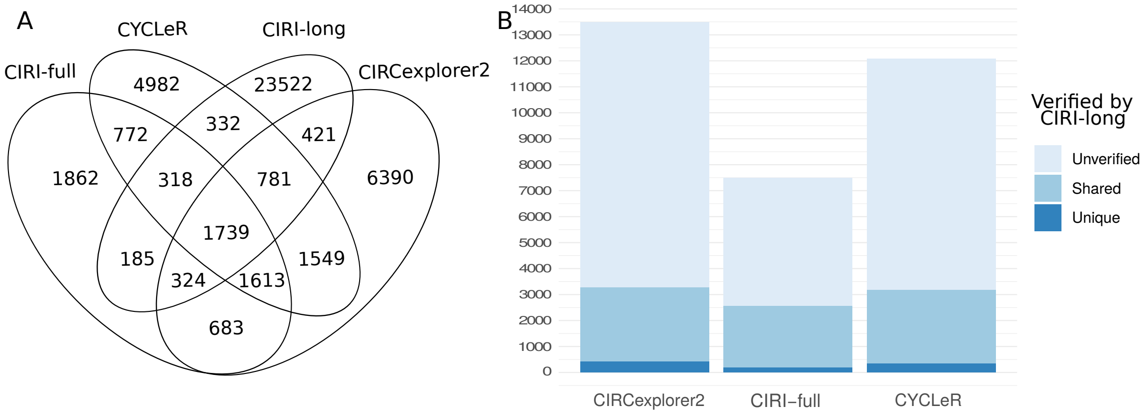


Figure 3.13: **CIRI-long Nanopore benchmark.** *These figures show the results of the comparison between Illumina-based methods and an Oxford Nanopore-based method. (A) shows a Venn diagram of the full set of assembled transcripts for each tool. (B) is a bar graph representation of the same data, but with emphasis on overlapping regions from the Venn diagram. On (B), the assembled transcripts for each Illumina-based tool are divided into* verified *by CIRI-long or* unverified*. The latter are further subdivided into* unique *- the transcripts that are shared only by one Illumina-based tool and CIRI-long, and* shared *- the transcripts that are shared by two or more Illumina-based tools and CIRI-long. CIRI-full has the lowest transcript count in every category. This is due to the length limit of assembly based on the library insert size. When comparing CIRCexplorer2 and* CYCLeR *, we notice that CIRCexplorer2 has only ∼100 more* verified *transcripts, while simultaneously having ∼1400 more unverified transcripts. Based on the information provided by the simulated benchmark, it is safe to conclude that the additional isoforms produced by CIRCexplorer2 are primarily erroneous assemblies.* ***Reproduced from Stefanov et al. (2022)***

good balance between precision and recall.

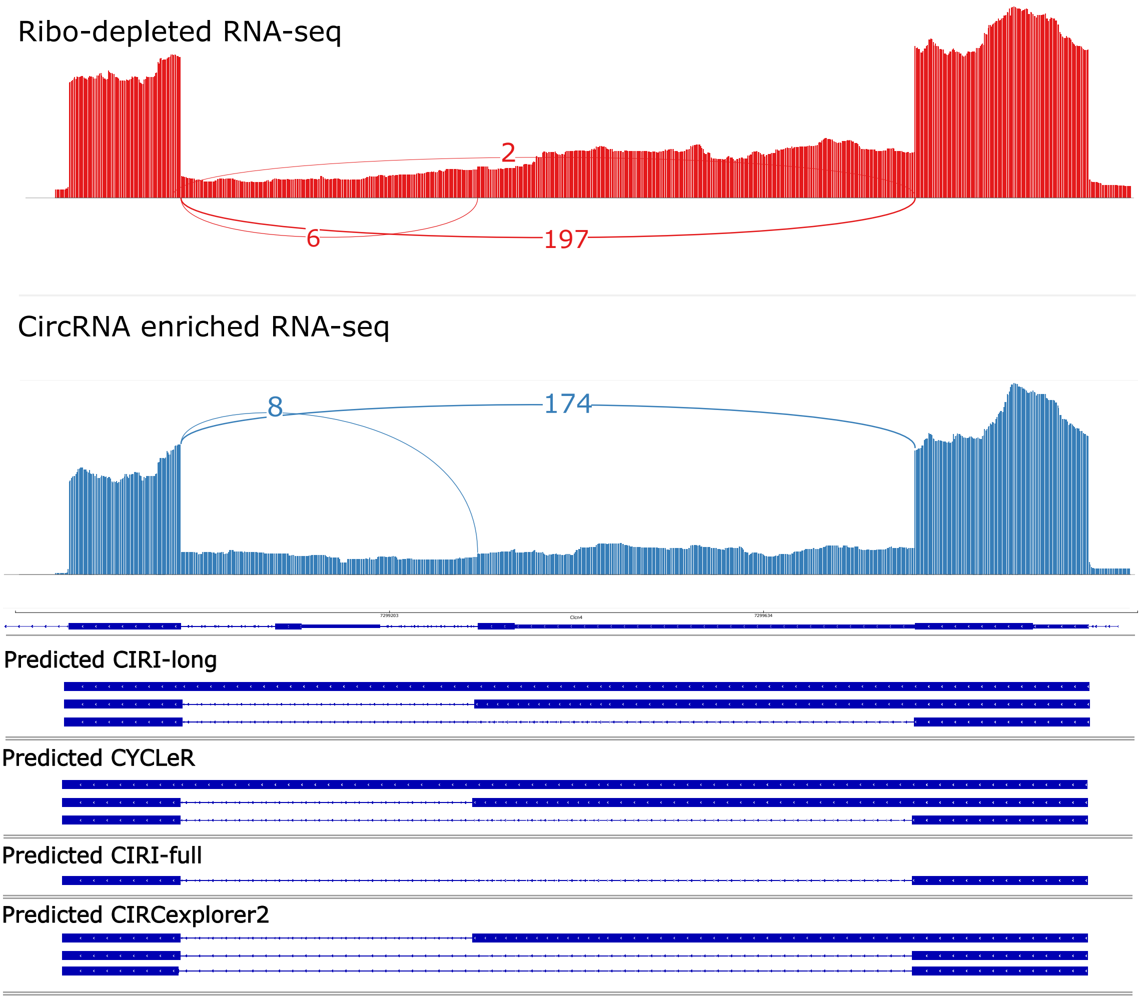


Figure 3.14: **Comparison of the assembly of BSJ locus chr7:7,298,969-7,299,877 in *M. musculus*.** *The Sashimi plots show STAR mapping of the exons and FSJ encompassed by the BSJ sites of the chr7:7,298,969-7,299,877 circles. The FSJs further away not participating in circRNAs have been removed for visibility. The plot shows a comparison between total ribo-depleted RNA-seq and circRNA enriched RNA-seq. The CIRI-long output serves as a true positive reference. This example was selected to show the advantage of* CYCLeR *to handle retained introns. As shown, CIRI-full does not account for an alternative 5'-splicing as well as the retained intron. CIRCexplorer2 makes an assembly error as the tool attempts to match the assembly to the given linear annotaion. The superior feature selection of* CYCLeR *compared to CIRCexplorer2 is the reason for avoiding an exclusively linear transcript FSJ, thereby preventing an incorrect isoform assembly.* CYCLeR *is the only tool that manages to assemble the isoform containing the retained intron.* ***Reproduced from Stefanov et al. (2022)***

3.6.4 Analysis of *D.* melanogaster data

The common approach applied in most circRNA studies is to make inferences based solely on information on the number of reads spanning the BSJs. An alternative approach is presented by sailfish-cir, by means of quantification based on a model of putative circRNA transcript. This comparative study is meant as a proof-of-concept of the workflow of *CYCLeR*. It tests the idea that assembly based on only a few key time points is sufficient to provide better transcript quantification and, therefore, better functional inference. The difference in quantitative output requires different normalisation strategies. In the case of CIRCexplorer 2, I applied CPM, while for sailfish-cir and *CYCLeR* I used RPKM. Subsequently, all counts are subjected to variance stabilisation through the use of the VST method from the DESeq2 package [60]. It is essential to acknowledge the number of overall BSJs per tool included in this analysis, as detailed in Table 3.7. The results shown for CIRCexplorer 2 include statistics from all detected BSJ sites. In the case of sailfish-cir BSJ sites are excluded based on the linear annotation. *CYCLeR* processes only the BSJ sites that can be identified in the matching RNase R data.

	BSJs	Transcripts
CIRCexplorer2	12554	-
sailfish-cir	11117	11515
CYCLeR	4371	5659

Table 3.7: D.* melanogaster data set: total number of identified transcripts.** *Summary of the full number of BSJs and transcripts that have been identified by the corresponding tools.* ***Reproduced from Stefanov et al. (2022)

The primary focus of this study is to emphasise the need for quantification of the full transcript, as opposed to focusing solely on reads mapping to BSJs. As we can see in Figure 3.15 (A), the reads spanning the BSJ are sufficient to distinguish between the general cell type and stages of development. When comparing Figures 3.15 (A) and (B), we see that the output of *CYCLeR* does not result in the loss of functional information, even with a lower number of annotated BSJs. Thus, a finer comparison between the qualification of CIRCexplorer2 and *CYCLeR* is required. Figures 3.15 (C) and (D) focus only on samples coming from *D.* melanogaster embryo. The results are presented as a dendrogram, where we can see a clear difference in the pattern of sample similarity. The similarity of the samples in the *CYCLeR* dendrogram matches the expected similarity between the developmental stages. The results of *CYCLeR* are more biologically relevant, as there we can see a clear separation of the circRNA expression pattern between samples of pre-14th hour and past-14th hour of development. This change is related to the

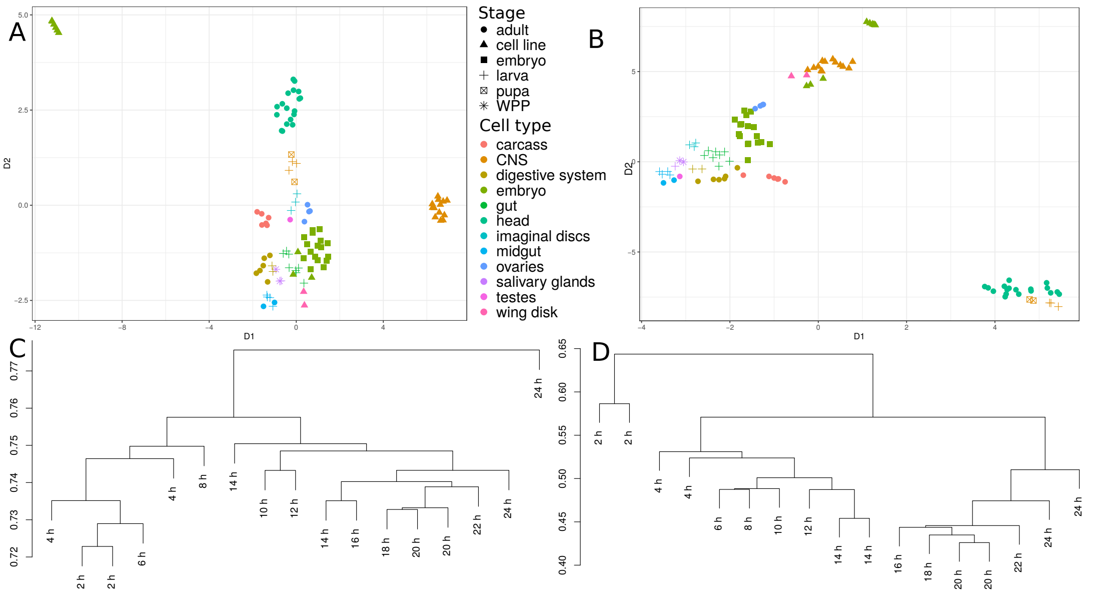


Figure 3.15:* Comparison of CIRCexplorer2 and *CYCLeR* for the *D.* melanogaster transcriptome sets.** *The comparison is made based exclusively on circRNA abundance estimations. (A) and (B) show the UMAP dimensional scaling of the abundances inferred by CIRCexplorer2's BSJ detection module and* CYCLeR *for all 103 samples of the Lai dataset. (C) (CIRCexplorer2) and (D) (*CYCLeR*) show a dendrogram of the subset of data corresponding to embryo stages which are based on between sample distance calculations. The scale of the dendrograms represents the samples' distances.* ***Reproduced from Stefanov et al. (2022)

expression of the *mbl* gene, which is a splicing factor with major effect on the expression pattern of circRNAs in *D.* melanogaster [9]. Equivalent graphs for the results of sailfish-cir can be found in Figure 7.1.

The samples of embryo stages are collected in different batches, which shows an advantage of *CYCLeR*. In Figure 3.16, we can see the UMAP scaling only of the results of the embryo samples, as well as annotation of the batches based on SRA accession numbers. We can observe notable differences between results from *CYCLeR* and the competitive tools. When using *CYCLeR* for quantification, it is possible to identify a gradient in the data that reproduces the known developmental stages. As a result, the outliers in the data become easily distinguishable. The results from *CYCLeR* are undeniably more biologically relevant than results based on reads spanning BSJs. This difference most likely stems from the fact that replicates have more stable variance when quantified with *CYCLeR*, shown in Figure 3.17. Comparison with sailfish-cir in Figure 3.16 allows us to highlight the advantage of the *de novo* assembly of the circRNA

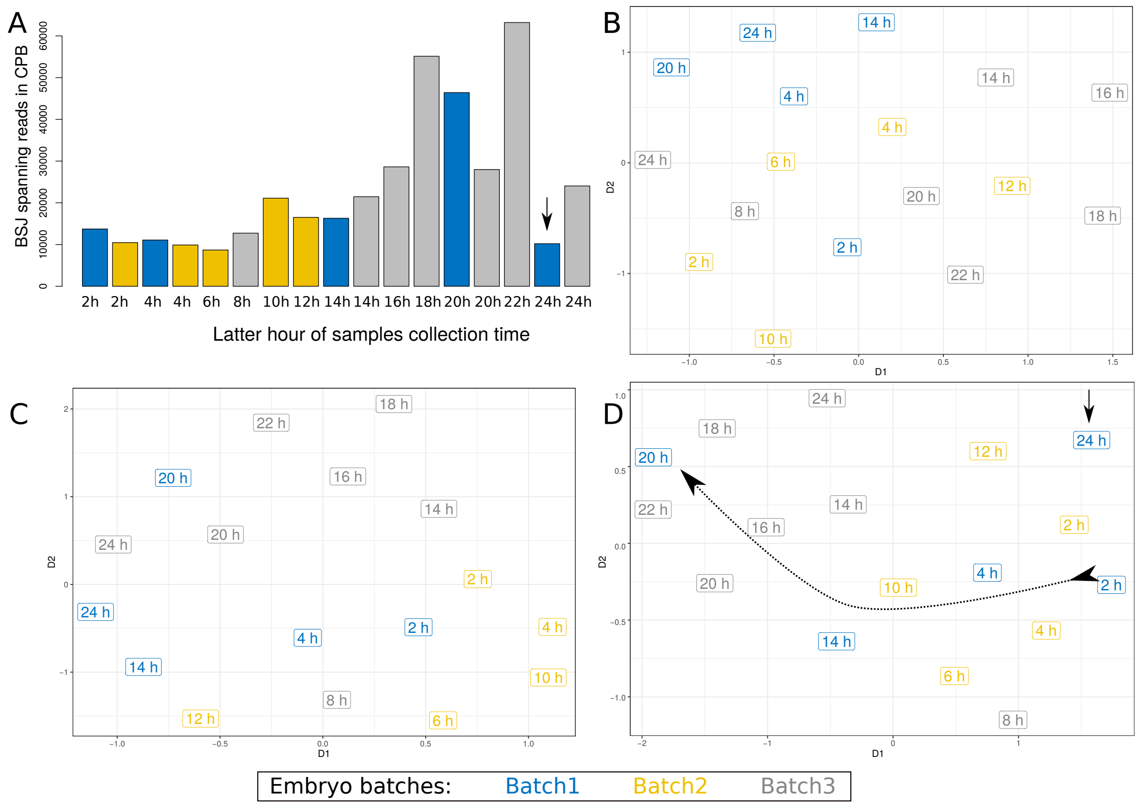


Figure 3.16: **Batch effect in Lai 2014 dataset.** *(A) shows the BSJ-spanning reads count per samples in counts per billion (CPB). (B), (C) and (D) show the UMAP dimensional scaling of the abundances estimated by CIRCexplorer2 BSJ detection module, sailfish-cir and* CYCLeR *respectively. I have annotated the experimental batches based on SRA accession numbers and colour-coded them. Only the* CYCLeR *results reflect the underlying biological trend in the distribution of the sample points indicated by the dotted curve. The trend is not perfect, due to the influence of strong experimental biases, but sufficient to reliably identify outliers (marked with straight arrow) and improve downstream analyses.* ***Reproduced from Stefanov et al. (2022)***

transcriptome and its influence on the quantification of the full isoform. In summary, the similarity measurements between the sample transcript abundances that are estimated with *CYCLeR* are significantly more biologically meaningful, emphasising that the correct assembly of isoforms and quantification based on the full transcript sequence is key to correct clustering of samples.

3.6.5 Benchmark with respect to qPCR values

As mentioned in 3.5.2.3, using qPCR values for full isoform assembly evaluation is challenging. The issue is exacerbated by the fact that publicly available results focus solely on two-exon circRNAs. The only dataset that fits the requirements of our study is the data used

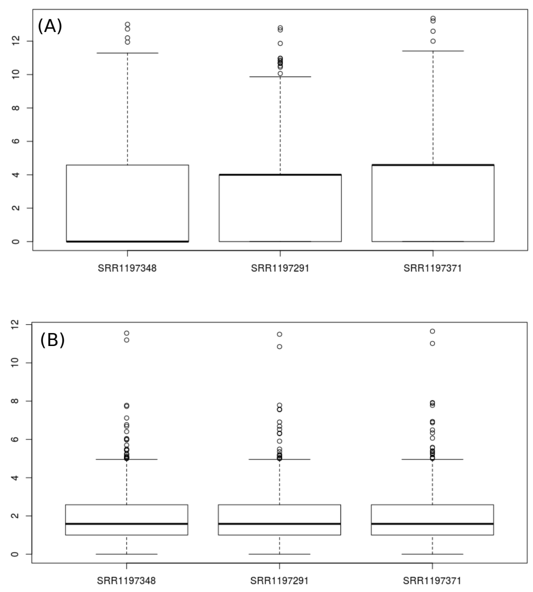


Figure 3.17:* Variance stability in third instar larvae, wandering stage, CNS samples.** *The median of the box plots of the normalised* CYCLeR *(B) counts is more stable than the normalised CIRCexplorer2 BSJ counts(A). This shows the higher variance stability between replicates, when quantifying with* CYCLeR. *The outputs are filtered to contain results from BSJ sites shared between the tools.* ***Reproduced from Stefanov et al. (2022)

in [118]. The study focuses on 13 BSJs, however, the BSJ locus of CAMSAP1 (Chr9:135881632-135883078) is known to have AS [50]. Additionally, *CYCLeR* identified alternative isoforms for the BSJ locus CORO1C (Chr12:108652271-108654410). This immediately highlights an issue with the currently available circRNA quantification tools – their inability to output results from circRNAs with alternative isoform. I am forced to eliminate results from those two BSJs from the study

When comparing the results from *CYCLeR* and the results of the quantification module of the CIRCexplorer pipeline (CLEAR), we can see only a minor discrepancy, leading to a correlation of 0.95. I evaluated two potential sources of the discrepancy – transcript length and GC-content difference between the full transcript sequence and the BSJ region. In Figure 3.18 (A) and (B), we see that the outliers of data match the divergence in GC-content, while transcript length seems to have no contribution to the difference.

The focus on the BSJ region, gives CLEAR an advantage over *CYCLeR* when comparing with qPCR values as seen in Figure 3.18

(C) and (D) – 0.75 vs 0.67. Given that the region assessed by CLEAR significantly overlaps with the region targeted by the qPCR primers, the GC-content difference is mitigated. It is important to

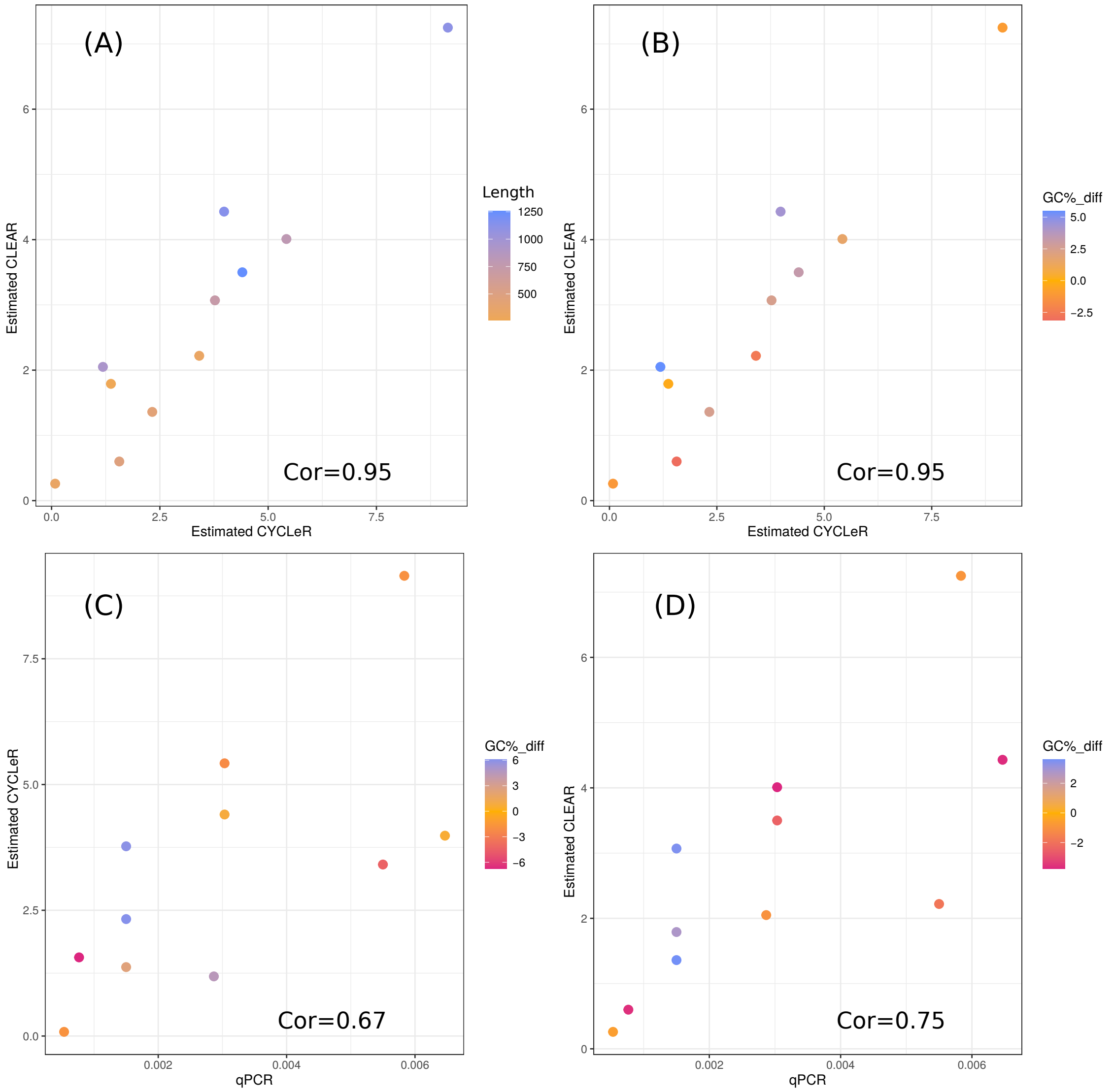


Figure 3.18: **Evaluation of the difference between CLEAR and *CYCLeR*.** *The output of* CYCLeR *and CLEAR are in very good agreement as shown in (A) and (B). The most likely sources of difference is the length of transcript (A) and GC-content (B). The comparison between the off-diagonal points in (A) and (B) indicates that the source of the difference is most likely the difference in GC content. The fact that CLEAR focuses only on the region around the BSJ makes the GC-content affecting CLEAR output closer to the GC-content affecting qPCR results. This is supported by (C) and (D) showing a comparison between the GC-content between the evaluated locus of the qPCR product to the GC-content of the* CYCLeR *transcript and 200 nt region around the BSJ respectively. The difference in GC-content is higher for the locus evaluated by* CYCLeR *and the off-diagonal points account for the difference between* CYCLeR *abundance estimation and qPCR results. Naturally, the differences between qPCR results and abundance estimation cannot be explained by those plots, as the difference between experimental procedures is influenced by a numerous biases, yet these plots at least manage to explain the better agreement between CLEAR and qPCR data.*

note that several experimental biases can lead to disparities between estimated qPCR and RNA-seq abundances, making it challenging for either tool to precisely align with qPCR results. However, the reduced performance of *CYCLeR* can be attributed to the fact that *CYCLeR* quantification is particularly sensitive to the differences in GC-content between PCR target region and full transcript.

The values used for the correlation calculation are summarised in Figure 7.5.

3.6.6 Consistency of assembly

To claim that *CYCLeR* provides robust output, it is necessary to showcase the consistent assembly between different samples of the same type. There are two important statistics that need to be evaluated: the consistency of assembly between replicates; and the consistency of assembly between samples that have undergone different enrichment procedures.

The previously analyzed *D.* melanogaster dataset has multiple libraries with replicates, as well as data from different generations. The publicly available data sets that provide libraries from the same source and with different types of circRNA enrichment are very limited. Specifically, just a single pair of libraries suits my needs. The PA1 cell line dataset, used in the quantification benchmark, has libraries with polyA-depletion treatment and a combination of polyA-depletion and RNase R treatments.

In Table 3.8, we can see a summary of the overlaps of assembled transcripts per sample. Biological replicates differ only by ∼10%. As expected, when comparing the transcripts between different generations, the difference increases. It is more challenging to compare the reconstruction based on libraries that have undergone different treatments for circRNA enrichment. Naturally, two types of depletion steps increase the number of detected BSJs compared to a single type of depletion – ∼8000 versus ∼34 500. To avoid any conflict arising from the difference in the starting BSJs set, I limited the starting set of BSJs to the BSJs derived from the total RNA-seq of the PA1 cell line without circRNA enrichment (∼2 500). This is a logical filter, because qualification is performed only on the untreated samples. The difference between samples with different treatments is close to the difference between biological replicates (see

Table 3.8). Naturally, this is insufficient information to draw general conclusions, but the results indicate that a single type of circRNA enrichment treatment is sufficient for practical purposes.

A summary of the transcript length and exon number of the predicted transcripts can be seen in Figure 7.4.

Sample A & Sample B	A\B	A∩B	B\A
D. melanogaster(Head) RNase R WT19: Rep1 & Rep2	311	3017	298
D. melanogaster(Head) RNase R WT28: Rep1 & Rep2	199	2343	196
D. melanogaster(Head) RNase R: WT19 & WT28	1889	1737	1001
PA1 cell line: PolyA(-) & PolyA(-)/RNase R	1003	6075	761

Table 3.8: **Summary of transcript assembly between different transcriptome samples.** *Pair-wise set difference and set overlaps between samples. Column 1 provides information on the 2 samples in the pair-wise comparison of reconstructed transcripts, columns 2 and 4 specify information about the number of different transcripts between samples and column 3 contains the number of overlapping transcripts.* ***Reproduced from Stefanov et al. (2022)***

3.7 Discussion

In this chapter, I have demonstrated the design, workflow, and advantages of the novel tool *CYCLeR*. After a thorough benchmark composed of multiple steps, it is very clear that *CYCLeR* outperforms the competition with regard to both transcript assembly and transcript quantification. This conclusion is supported by both simulated and real data. I separate the simulated in two datasets, mimicking respectively the standard type of reconstructions problems as well as more convoluted, rarely occurring cases. *CYCLeR* outperforms in assembly and quantification metrics for both dataset categories. The results from the simulated data are supplemented by verification based on long-read experiment. The results of the two studies are in very good agreement and highlight the major advantages of *CYCLeR* over the competitive tools.

One key advantage of the *CYCLeR* assembly module is the robust feature selection, based on a statistical test for replicates. Although there is a great overlap between the results of the competitive tools, the true positive results that are unique to *CYCLeR* are those that involve novel transcript features. Although it is commonly disputed whether most circRNA even have functions, circRNAs that contain transcript features that are not present in linear transcripts are considered prime candidates for further studies. Even if *CYCLeR* was not outperforming in all metrics, this feature of the tool alone would be sufficient to justify its use. Another key advantage of *CYCLeR* is the stepwise transcript reconstruction. This algorithm allows for the high precision of *CYCLeR* without sacrifice of sensitivity.

Neither of the two strategies in *CYCLeR* lead to immaculate results. Nevertheless, the pipeline design partially compensates for the reconstruction of false positive transcripts. Unlike competitive tools that perform quantification and assembly in parallel, *CYCLeR* separates the quantification step from the assembly. In fact, the strategy for quantification in *CYCLeR* does not rely on mapping of split reads, making it impervious to common mapping artefacts. False positives reconstructions, caused by mapping artefacts, will correspond to negligible relative abundances. Thus, the abundance estimation can serve as a filter for false positive transcripts. A comprehensive threshold for such filter is impossible to assign, because the transcriptomes of different organisms vary greatly between species in terms of AS occurrence, as well as exon size and number. Varying library depths also affect the value of a potential abundance threshold.

CYCLeR does not have explicit limitation on the library depth or insert size like the tools from class IV. All circRNA-specific analyses require a matching set of ribo-depleted and circRNA-enriched libraries. Often the ribo-depleted library is used just for the study of linear RNA while the circRNA-enriched is used exclusively for circRNA study. Subsequent merging of the data from experiments with different biases leads to inconsistencies in the study. *CYCLeR* is the only tool that makes full use of the data to perform assembly and quantification. One could argue that the tools from class IIIa use the entire data of ribo-depleted and circRNA-enriched libraries for an adjustment of the circRNA levels. However, the results produced by *CYCLeR* have better efficiency in using all data, as *CYCLeR* provides a unified output with circular and linear RNA quantities, without the need for additional adjustments.

A unique requirement for *CYCLeR* is the need for replicate libraries. However, I have shown with study of *D.* melanogaster dataset that circRNA enrichment of a few time points is sufficient to produce valuable results. Using the same dataset, I proved that using *CYCLeR* is essential for the correct clustering based on the quantities of the circRNA transcripts. In fact, the output of *CYCLeR* is more convenient to use as it provides combined abundance estimation of linear and circular transcripts in a format commonly used for downstream pipelines. This is an important design feature of *CYCLeR* that facilitates an inevitable future transition to single-cell non-coding RNA studies.

Chapter 4

Identification of functional circRNA-RNA interactions

This chapter covers the work that the doctoral candidate did, under the oversight of prof. Irmtraud Meyer (MDC-Berlin), to identify candidate circRNAs that participate in functional RNA-RNA interactions related to pluripotency maintenance. This work was done in collaboration with the experimental lab of Dr. Zsuzsanna Izsvak (MDC). I conceptualised the study's experimental design, aiming to integrate newly generated data with publicly available datasets. I have determined the protocols and navigated the parameters to be used to the generation of the RNA-seq libraries. I had minimal influence over the protocols related to transfection, cell selection, retinoic acid treatment and nuclear extraction. The novel RNA-seq data was generated by Dr. Aleksandra Kondrashkina and Dr. Cristine Römer. Subsequently, I devised and performed all computational analyses and data visualisation described in the chapter. At the time of the submission of the thesis, no data has yet been made publicly available and no results have been published.
Note: The use of ”we” throughout the text refers to the author-reader collective.

4.1 Insight into the experimental design

4.1.1 RNA-RNA interaction detection issues

RNA-RNA interactions (RRI) are a common away of the organism to regulate RNA processing. As explained in Section 2.5, the computational search for RRIs based on transcript sequence can lead to many false positive results. However, the addition of information on evolutionary conservation of the RNA duplex increases the likelihood of a putative duplex being a true positive [92, 94, 93]. Evolutionary conservation is also a sign that a duplex is functionally significant. Nevertheless, even with the addition of evolutionary information, the pure computational search of RRI is unreliable and needs to be supplemented by experimental data.

4.1.2 Co-expression network

Co-expression networks has been used to identify molecule interactions for decades [64, 130]. Although the approach usually focuses on gene co-expression networks and the discovery of novel protein-protein interactions, the strategy is readily applicable to transcript co-expression and interactions, see Section 2.4.4.

The underlying assumption when studying co-expression networks is that genes (or trascripts) that participate in the same functional pathway are co-regulated. Therefore, transcripts that participate in functional interaction as a part of a particular biochemical pathway are likely to be co-expressed. Co-expression network analysis is based on a calculation of the correlation between relative abundances of the transcript. As a result, to achieve optimal calculations, an appropriate number of samples is needed. As a general rule, 12 samples are considered necessary to produce reliable correlation results. Naturally, samples should have a high variance in transcript levels to facilitate optimal information gain by correlation.

As I am interested in transcripts that have an effect on pluripotency, I chose to focus on the differentiation of hESCs. I design my study with a focus on pluripotency because there is an overabundance of publicly available datasets that focus on studying different biochemical mechanism (e.g. transcription factor (TF) binding, protein-RNA interactions etc.)

My experimental design is based on the time series experiment of differentiation of the H1ESC cell line induced by treatment with retinoic acid. Treatment with retinoic acid (RA) forces pluripotent cells to exit the pluripotent state. When stem cells are derived from mice, RA treatment leads to a neuronal phenotype of cells [131], in human ESC the situation is more convoluted [132, 133, 134, 135]. RA treatment forces the hESCs to exit pluripotent state, but does not commit them to any specific type of differentiation. The addition of specific differentiation factors can commit cells to a specific fate [136, 137]. However, my primary focus are transcripts related to pluripotency and I do not need the cells to commit to a specific cell type. In fact, the undefined state is likely to cause more variance of transcript levels after RA treatment, which is beneficial for correlation calculations. After the fifth day of treatment, cells start to show signs of necrosis, and by the seventh day the cell population decreases significantly. This means that a reasonable time to end the time-series experiment is on the fifth day. To ensure sufficient variance between samples, batches of cells are harvested daily. Replicates of each stage are needed to ensure that the study is not influenced by outliers. Also, the presence of a replicate for each stage gives the option to perform a ”rudimentary” differential expression analysis between stages.

To expand the number of time points, I added another stage to the experimental design, which I will term HERVH-high cells. HERVH is a retroviral transposon, and transcripts that result from the integration of these repetitive elements are known to be specifically expressed in pluripotent cells [138]. HERVH-high cells are a selected population of hESC, in which the levels of TF binding to HERVH/LRT7-derived sequence is higher. Based on a GFP reporter construct with a LRT7 and HERVH-int sequence, single clones of hESC cells can be selected and used for growth of a new line. HERVH-high cells have been shown to exhibit expression patters with very high similarity to naïve ESC, making them perfect for studying pluripotency mechanisms [139].

The replicates of the time series experiment are intended to come from separate batches. By implementing such an approach, despite samples being replicates, they can have variances that would help follow-up clustering based on co-expression of transcripts. There is an additional purely practical advantage – ensuring that the RNA-seq libraries from the first batch have the parameters that I require to avoid wasting resources on sub-optimally generated data.

4.1.3 RNA proximity ligation data

4.1.3.1 SPLASH

While co-expression networks indicate potential functional association between transcripts, they are not a direct proof of interaction. For a direct proof of an interaction, a high-throughput protocol is needed for the examination of in vivo RNA duplexes, see Section 2.5.3. There is publicly available data that supplement the design of the time-series experiment. SPLASH is an RNA cross-linking and proximity ligation protocol capable of the detection of RNA duplexes in vivo [107]. The details of the experimental procedure are shown in Figure 4.1. An important feature of the experimental protocol is the polyA-enrichment procedure that limits the circRNA RRI search to circRNAs that interact with transcripts that have a polyA-tail. Furthermore, circRNAs are known to be more resistant to fragmentation, see Section 2.3.2.2. This means that we can expect direct duplex detection only from circRNAs with considerable length.

The proximity ligated fragments are detected as chimeric fragments in the final SPLASH library. Importantly, non-chimeric reads are not useless, as they also provide information about occurring RNA duplexes. The corresponding RNA fragments are enriched by the RNA duplex selection, therefore, they can be used in enrichment analyses, see Section 4.4.2.1. In the SPLASH protocol, the transcripts are fragmented for the sake of precise proximity ligation, indicating the loci involved in a duplex. However, the small size of the SPLASH fragment becomes a disadvanatage when attempting to assign the duplex to a specific isoform, see Figure 4.2. Unless the chimeric read also spans a splice junction or an exon specific to a single isoform, it is a near impossible task to definitively assign a duplex to an isoform. An advantage of the SPLASH data as opposed to other RNA proximity ligation approaches is the fact that fragment sizes produced by the experiment are relatively long (80-120 nt) and have a decent chance of mapping both as a duplex chimera and a splice junction site.

4.1.3.2 CLIP-seq

The first circRNAs with identified function is circCDR1as. This circRNA has functions as a potent microRNA sponge. As a turning point in the study of circRNA, this discovery influences the study of circRNA to this day. It has become a common practice to test the potential of any circRNA of interest to be a miRNA sponge. As explained in Section 2.5.3.2, AGO-CLIP is the optimal high-throughput approach to detect miRNA-RNA interactions. With

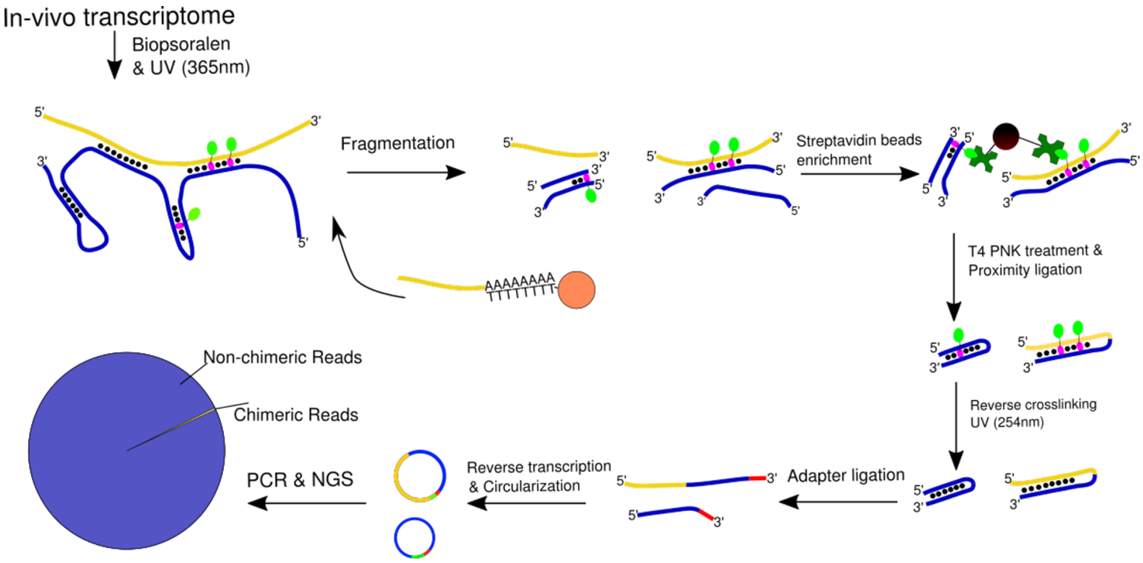

In-vivo transcriptome
Biopsoralen & UV (365nm)
Fragmentation
Streptavidin beads enrichment
T4 PNK treatment & Proximity ligation
Reverse crosslinking UV (254nm)
Adapter ligation
Reverse transcription & Circularization
PCR & NGS
Non-chimeric Reads
Chimeric Reads
AAAAAAAA
5'
3'

Figure 4.1: **Detailed workflow of the SPLASH protocol** *The first step of the standard SPLASH protocol is the intercalation of biotinated psoralen (biopsoralen) compound in an RNA duplex. The biopsoralen is then crosslinked with 365 nm UV light. The RNA is fragmented by thermal fragmentation in the presence of Mg ions. The duplexes with biopsoralen integration are then pulled-down with streptavidin beads. The RNA strands of the duplexes are treated with T4 PNK to prep the RNA ends for ligation. The 2 strands of the duplex are then ligated. The biopsoralen crosslinking is reversed with 254 nm UV light treatment. The resulting RNA fragments undergo adapter ligation, circularisation and reverse transcription. The following fragment undergoes standard RNA-seq library preparation. Only chimeric reads hold direct information for RNA duplexes. However, the resulting library has only 1% chimeric reads due to inefficiency of the protocol. An additional polyA-enrichment step can be added prior to the fragmentation procedure. Figure adapted from [140], available under CC BY 4.0.*

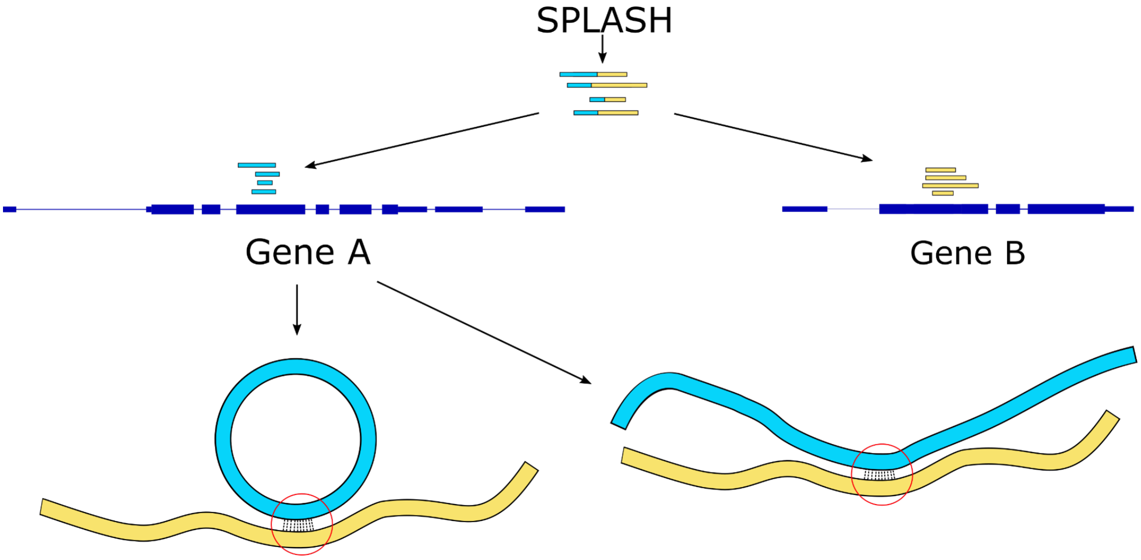


Figure 4.2: **Challenges with the use of SPLASH data for a study with a focus on a specific isoform** *The mapping of the chimeric read reveals only a small portion of the sequence of interacting isoforms. Purely from the chimeric read, it is impossible to know which isoform of gene A interacts with a transcript of gene B.*

my focus on pluripotency, I benefit from publicly available data for H1ESC AGO2 CLIP-seq [141].

4.1.4 CircRNA enrichemnt

For circRNA assembly, CYCLeR requires pairs of ribo-depleted and circRNA-enriched RNA-seq libraries. I have shown in Section 3.6.4 that the information from a few key time points is completely sufficient for the functional analysis of circRNAs. I have selected H1ESC, H1ESC on day 5 after RA treatment, and HERVH-high cell as the most relevant time points. Naturally, circRNAs that are specific for a very particular time point of differentiation will be lost in my analysis, but as previously mentioned, they are not the primary focus of my study.

4.1.5 Nuclear enrichment

RNA processing is specific for the nucleus and often occurs co-transcriptionally. This means that for a circRNA to be involved in RNA processing, it needs to be at least partially present in the nucleus. To identify circRNAs enriched in the nucleus, my experimental design involves nuclear-fraction-specific RNA sequencing from the steps of untreated H1ESC samples, day5 of RA treated samples and HERVH-high samples, based on the same logic used in Section 4.1.4. We conducted a trial experiment to determine whether there is a necessity to reconstruct circRNAs from a nuclear fraction and create corresponding circRNA-enriched RNA-seq libraries. The results showed that the assembly of circRNAs from the nuclear fraction is not only redundant, but misleading. This is due to the high level of nascent RNA that mimics intron retention.

4.2 Data collection

4.2.1 RNA-seq library preparation

The key step in this study is transcriptome assembly and quantification. It is generally advised to prepare RNA-seq libraries with longer insert size to facilitate transcript reconstruction. However, due to the increased stability and smaller sizes of the circRNAs (see Section 2.3.2.2), the RNA-seq libraries for my study require more stringent fragmentation conditions. I have determined the optimal median fragment size of the libraries to be 190 nucleotides, achieved through RNA Mg^{2+} ion fragmentation.

CYCLeR requires circRNA-enriched libraries as input. In my experimental design, RNase R enrichment is used as a circRNA enrichment strategy.

The data was generated by Aleksandra Kondrashkina and Christine Römer from the lab of Zsuzsanna Izsvák. A summary of the set of RNA-seq libraries can be seen in Table 4.1.

Sample type	Fraction	RNase R Treatment
HERVH-high	nuclear fraction	no
HERVH-high	whole cell	yes
HERVH-high	whole cell	no
HERVH-high	nuclear fraction	no
HERVH-high	whole cell	yes
HERVH-high	whole cell	no
ESC	whole cell	no
ESC	whole cell	yes
ESC	nuclear fraction	no
ESC	nuclear fraction	yes
ESC	whole cell	no
ESC	whole cell	yes
ESC	nuclear fraction	no
ESC	nuclear fraction	yes
RA day1	whole cell	no
RA day3	whole cell	no
RA day4	whole cell	no
RA day5	nuclear fraction	no
RA day5	whole cell	yes
RA day5	whole cell	no
RA day1	whole cell	no
RA day3	whole cell	no
RA day4	whole cell	no
RA day5	nuclear fraction	no
RA day5	whole cell	yes
RA day5	whole cell	no

Table 4.1: **Summary of the generated RNA-seq libraries** *The tag RA day indicates the day after the start of RA treatment.*

4.2.1.1 HERVH-high sample preparation

Transfection and sorting of H1ESC was done according to the procedure in [139]. Two separate single-clone lines were used for our study.

4.2.1.2 Time series sample preparation

In preparation for the experiment 90% confluent H1ESC (WA01 alias, 18-W0260, WiCell), cultured in full mTeSR 1 media (85870, StemCell Technologies) with added antibiotic primocin (ant-pm-05, Invivogen), on vitronectin (A14700, Life Technologies) coating. Cells were split in clumps with versene (15040066, Thermo Fisher Scientific), 1:10 ratio, to have 15 plates (90 wells) (11337694, Thermo Fisher Scientific). Rock inhibitor (Y-27632, StemCell Technologies) was added to inhibit apoptosis of cells after splitting. Treatment was done with 10µM retinoic acid (R2625 Sigma) in mTESR media. The media was changed daily. Cell were collected daily for the following five days.

4.2.1.3 RNA-seq library preparation

Fractionation was done while collecting cells with Nuclei EZ lysis buffer (N3408, Sigma). Nuclear or whole cell pellets were frozen in Trizol (15596026, Thermo Fisher Scientific). After collecting all samples, total RNA was isolated simultaneously with Quick-RNA Miniprep Kit (R1055, Zymo Research) kit. CircRNA enrichment was performed with Ribonuclease R (RNase R) (RNR07250, Lucigen epicentre): 1.5ug RNA were digested with 10U of RNAse R for 30min at 37°C. Then RNA was purified in the equal volume of phenol chloroform isoamyl alcohol, the upper aquatic phase was the purified with 4M LiCl.

For library preparation, Roche KAPA RNA Hyper+RiboErase HMR was used according to instructions. Libraries were prepared from 500ng RNA, except for nuclear fraction day 5 RA – 230ng. RNAse R treated sample was used the maximum possible – 10uL. Temperature fragmentation was performed for 7 min; except for RNase R treated samples – 6 min, and the nuclear fraction sample of day 5 of RA treatment. The HERVH-high libraries have undergone 5 cycles of PCR amplification, while the time series libraries–8.

4.2.2 SPLASH data

The sequencing data is already publicly available under the SRA accession ID SRP073550. The H1ESC replicates are stored as

SRR3404926 and SRR3404943, while the H1ESC treated with RA are stored as SRR3404927 and SRR3404928 [107]. The archived data is processed with fastq-dump from the SRA Toolkit.

4.2.3 CLIP-seq data

AGO2 PAR-CLIP-seq data for H1ESC are available under SRA accession ID SRR359787 [141]. The archived data is processed with fastq-dump from the SRA Toolkit.

4.3 Quality control of the novel data

We have performed a trial run of library preparation and sequencing (data not shown). After the trial run, all library parameters were adjusted to fit my desired specifications. There was a single troubling statistic after quality control check on the FASTQ files with FastQC [142] – high duplicate rate.

The initial quality control of the libraries indicates alarming rates of duplicated reads – reaching up to 50-60 % duplicates. Naturally, mapping of the pair-end reads decreases the number of duplicated read statistics, but further study of the cause the high number of reported duplicates is required to exercise caution.

Usually, high percentage of duplicates is an indicator of a high number of PCR cycles during the library preparation protocol. However, reads that map to multiple loci are indistinguishable from duplicates. Thus, we could not determine how much of the detected duplicates are due to technical reasons (PCR duplicates/Optical duplicates) and how much due to natural repetition of library fragment sequences. It is important to note that the smaller insert size of the library increases the chance of multi-mappped fragments, specifically for fragments originating from transcripts with repetitive element integration. Furthermore, the transcriptome mapping tools [54, 55] (kallisto, salmon) do not deal with duplicated reads, which means that RNA-seq data would need to be preprocessed to remove duplicates.

4.3.1 Common solutions and tools

- samtools [124] - collapses the reads/fragments with identical sequence to a single read

- picard [143]– marks the reads with identical sequences without removing the repetitions; the user can subsequently select whether to count the marked reads
- STAR [34] - marks the reads with identical sequences without removing the repetitions, but also has the option to not mark the multi-mapping reads

4.3.2 Analytical strategy

Given the high likelihood of reads coming from repetitive regions, it is very likely that a portion of the reads that were marked as duplicates by the initial quality control are actually multi-mapping reads. To resolve that, I used the STAR option to mark duplicates, but ignore regions of multi-mappings.

featureCounts [125] counting of the BAM files with genes as features and different parameters regarding the counting of duplicate reads or multi-mapping reads. Note that chimeric reads (circRNA) are also considered multi-mapping. The counting statistics allow us to see how much of the reads are technical duplicates, what the influence of multi-mapping reads on the data is, and how much of the reads are coming from unannotated transcripts. Important to note is that multi-mapping reads are not fractionally counted. I have performed four separate counting procedures: all reads counted, counting without marked duplicates, counting without multi-maps, counting without both multi-maps and marked duplicates.

4.3.3 Results and Conclusions

Table 4.2 shows the percentage of reads/fragments counted based on the hg38 Ensembl annotation, see Table 7.2. Some trends are noticeable. When counting the only without marked duplicates, there is only a sight drop in the percentage mapped reads. It is important to note that this percentage is not fully representative due to the fact that there are naturally some PCR duplicates among the multi-mapped reads. But most importantly, the change that not counting multi-mapped reads far exceeds the number when duplicated reads are not counted, explaining the high number of fragments with the same sequence, and thus explaining the alarming number of PCR duplicates in the FastQC results. Note that reads

Sample type	F	RR	% all	% no dup	% no multi	% no multi/dup
HERVHhigh	nf	no	73.7	71.9	52.7	50.3
HERVHhigh	wc	yes	59.8	56.7	19.8	16
HERVHhigh	wc	no	55.2	54.2	19.1	17.2
HERVHhigh	nf	no	60.9	60	31.9	29.8
HERVHhigh	wc	yes	46.5	44.4	1.2	1.1
HERVHhigh	wc	no	81.8	72.1	61.6	51.5
RA day1	wc	no	72.9	56.5	45.2	28.2
RA day3	wc	no	83.5	59	58.9	34
RA day4	wc	no	55.3	50.7	14	8.3
RA day5	nf	no	79.8	59.2	62.9	42.1
RA day5	wc	yes	43.3	43.3	2.7	1.6
RA day5	wc	no	85	59.1	66	39.9
ESC	wc	no	82.1	71.8	61.5	50.9
ESC	wc	yes	60.2	55.3	25.6	20
ESC	nf	no	72.7	66.7	48.7	42.3
ESC	nf	yes	57.1	54.7	15.1	11.8
ESC	wc	no	80	70.4	57.8	47.7
ESC	wc	yes	58.9	54.4	21.4	16
ESC	nf	no	68.6	63.7	42.9	37.5
ESC	nf	yes	49.2	48.5	10.5	8.7
RA day5	nf	no	80.5	62.6	60.8	42.6
RA day5	wc	yes	63.9	57.4	18	10.6
RA day1	wc	no	57.3	49.6	25.9	17.3
RA day3	wc	no	66.9	56	33.3	21.7
RA day4	wc	no	73.5	57.3	45.2	28.4
RA day5	wc	no	76.9	59.6	52.7	34.9

Table 4.2: **Summary of multi-mapping reads vs duplicates** *The table shows the percentage of counted reads per sample based on STAR duplicate and/or multi-map marking. RR stands for RNase R treatment. Nuclear fraction (nf) and whole cell (wc) are in the fraction column (F). The results are separated in counting all reads (% all), counting all read except those marked as duplicates (% no dup), counting all read except those marked as multi-mapped (% no multi), and counting all read except those marked as duplicates or multi-mapped (% no multi/dup). The difference between number of mapped read when comparing counting of all possible reads and when counting with skipping marked reads, explains the high number identical reads. When eliminating multi-mapped reads from the counting the difference is substantially higher than when eliminating reads marked only as duplicates. Also, samples with highly active endogenous retroviral elements are more affected by the multi-map filter. Therefore, the identical sequences are mostly caused by fragments originating from repetitive element loci or circRNAs.*

mapping to circRNAs will also be considered multi-mapping. Therefore, the percentage of mapped reads, when multi-maps are ignored, decreases in the RNase R treated samples. Libraries generated from a higher amount of starting RNA have a lower (disregardable) percentage of true duplicate reads from PCR. Duplicate rates less than 20% are generally ignored because the computational removal of duplicates between samples can cause skews in the data. Therefore, even for samples that could theoretically benefit from duplicate removal, it is not advisable to remove them, as they will cause discrepancies compared to other samples. The percentage of reads mapping to annotation decreases for libraries generated from nuclear fraction or RNase R treated samples. Some of those reads belong to circRNAs, but it is likely that there are a high number of unannotated linear transcripts. The number of reads mapping to unannotated regions increases with pluripotency decrease. This means that there are many yet unannotated transcripts related to pluripotency, partially originating from active endogenous retroviral elements [144].

4.4 Computational workflow

4.4.1 BSJ detection

The BSJ detection for all RNA-seq samples was carried out in a way that satisfies the requirements of CYCLeR, see Section 3.3.1. For the downstream analysis, BSJ detection was performed for all samples of the SPLASH data. However, tools for BSJ detection cannot distinguish between chimeric reads that are produced by RNA proximity ligation corresponding to the RNA structure and chimeric reads that correspond to a BSJ. I used the results from BSJ detection from RNA-seq samples to filter the results coming from SPLASH samples, keeping only the BSJ sites from SPLASH that match the BSJ sites in RNA-seq.

4.4.2 BSJ enrichment analysis

4.4.2.1 Prerequisite

The problem of assigning a specific isoform to an RNA duplex (shown in Figure 4.2) can be overcome to some extent. The SPLASH data for which I am interested have undergone a polyA-enrichment procedure. Theoretically, that means non-polyA transcript will be abundant in the data only if they interact with polyA transcripts. To verify this hypothesis, I performed gene-wide read counting on polyA-enriched SPLASH data for the cell line GM12892 (SAMN04870489, SAMN04870490, SAMN04870491, SAMN04870492) and matching polyA RNA-seq (SRR1803204, SRR521534, SRR521535 [145, 146]) with featureCounts. I used the resulting count matrix to perform a differential gene expression analysis with DESeq2 [60]. The results show a high enrichment of non-polyA genes in the SPLASH data compared to RNA-seq data. Specifically, the genes from the snRNA, snoRNA and rRNA biotypes are enriched in the SPLASH data. These are RNAs with a generally known function that requires base pairing. Therefore, if there are circRNAs that are enriched in SPLASH data, it is logical to assume that they are involved in RRI.

The proof of concept test is based on data that have multiple replicates with a library depth similar between replicates. In such a case, DESeq2 can be used safely to check for enrichment. However, the SPLASH data on ESCs have only two replicates with 10-fold difference in library depth. Therefore, an enrichment test in the ESC data requires an alternative strategy.

4.4.2.2 BSJ enrichment test

To perform an enrichment analysis for circRNAs, I devised an over-representation test to check the enrichment of BSJs. It is less sensitive than the DESeq2 alternative, but much more robust, in the sense that the resulting small number of enriched circRNAs is more reliable. BSJs can be used as a rough estimate of circRNA transcript levels. BSJ levels can also be compared. BSJ counts from a control library can be compared with BSJ counts from a treated library, and a hypergeometric test can be performed to calculate the probability of significant enrichment, see Section 2.4.5.1. I applied such an enrichment test to evaluate the enrichment of circRNAs in SPLASH data and later the enrichment of circRNAs in nuclear fraction.

4.4.3 CircRNA assembly and quantification

CircRNA assembly and quantification is performed with CYCLeR, following the procedure explained in Section 3.3.

4.4.4 BSJ enrichment

The enrichment analyses are based on the H1ESC dataset comprising control, nuclear extract library, and libraries treated with RNase R. The BSJ enrichment analysis manages to provide information of the most enriched BSJ after RNase R treatment; this is just a safety net that ensures that any BSJ result we have from the other test indeed corresponds to real circRNAs. The BSJ counts in the table are also thresholded by abundance (more than 5 reads supporting the junction). The selected BSJ sites are then used to test enrichment of BSJ for nuclear localisation and BSJ enrichment in the SPLASH data.

4.4.5 Co-expression analysis

4.4.5.1 Transcript selection

The count matrix with quantification based on CYCLeR is transformed with the variance stabilisation function of the DESeq2 package. To perform optimal clustering, only meaningful transcripts have to be selected; that is, transcripts with enough abundance and variance to provide a reliable correlation calculation, see Section 2.4.4. The selection of transcripts is done by exploratory data analysis. My workflow consists of iterative creation of tree diagrams and UMAP plots and visual observation of the distance between samples. Linear transcripts are selected on the basis of a variance cut-off. I used a generalised additive model fit of the ratio between the mean and variance of the transcripts across samples. The model is used as a reference point to select transcripts in which the variance-to-mean ratio deviates the most from the fit. I also ensured that the set of transcripts covers known markers for pluripotency. The same procedure could not be applied to circRNAs, as their expression levels are generally lower, making the fitting procedure unreliable. Thus, for circRNAs, I use a hard threshold on the overall abundance of circular transcripts across all samples.

The final set of transcripts selected for clustering consists of 10000 linear and 2600 circular transcripts.

4.4.5.2 Clustering

As a first step, I calculated an adjacency matrix for all transcripts based on a Pearson product correlation. I use the adjacency matrix as input for the calculation of the topological overlap matrix [64] and use the Dynamic Hybrid tree-cut algorithm [64] to separate the transcripts into clusters (for more information, see Section 2.4.4).

4.4.5.3 Gene set enrichment

The gene set analysis is based on genes and not transcripts. Therefore, every transcript is matched to a gene, and a set of unique genes is selected per cluster. For each cluster, I performed a gene ontology analysis using the topGO [147] package. For the clusters of interest, I performed an over-representation test using hypergeometric test based on the MsigDB sets for positional (C1), curated (C2), and regulatory target (C3) human gene sets, see Section 2.4.5.1.

4.4.6 SPLASH data analysis

Analysis of RNA proximity ligation data is limited by the steps of the computational pipeline, see Section 2.5.3.4. However, my goal does not require processing of the entire dataset. First, I am primarily interested in RRIs. Therefore, it would be redundant for me to process all the reads that correspond to RNA structures. Second, I limited the search space to the loci of a few very specific circRNAs selected based on result from the previous analytical steps. These 2 conditions allow me to design a very simplistic and robust analytical procedure to analyse the SPLASH data.

1. I mapped the reads from the SPLASH data with both STAR and BWA. The use of 2 mappers ensures that I am not unintentionally omitting chimeric read mappings, see Section 2.5.3.4.
2. I selected all IDs of reads that map to my loci of interest with samtools [124]. Afterwards, I extracted all reads that match those IDs. This step produces one SAM file per sample per locus.

3. I used all SAM files to generate a count matrix with featureCounts [125].
4. I observed which chimeras map outside the locus and were observed in multiple samples.
5. I used RIsearch [148] to identify the regions of the RRI with nucleotide resolution.

4.4.7 CLIP-seq analysis

The AGO2 PAR-CLIP-seq data is mapped with BWA [149]. The regions of interest are selected manually.

4.5 Follow-up computational procedures

4.5.1 Differential expression analysis

A common to approach to identifying the function of a transcript is the targeted decrease of its levels–Knockdown (KD). To evaluate the effect of *FIRRE* KD in mouse and human cells, I performed a differential expression analysis of the transcripts using data from the study with gaprmer KD of *FIRRE* transcripts in human and mouse cell lines with GEO-ID:GSE45157 [150]. RNA-seq libraries were quantified with kallisto using the standard Ensembl linear annotation for mm10 and hg38 as reference, see Table 7.2. The differential expression itself was performed using sleuth [58]. The resulting differentially expressed genes were used for topGO [147] gene ontology enrichment analysis.

4.5.2 Editome-based RNA folding

ADAR1 deaminates adenosine (A) to inosine (I) in cellular double-stranded RNA (dsRNA) substrates , thereby catalyzing the most common type of RNA editing found in humans. RNA A-to-I editing is performed to disrupt long RNA duplexes and leads to a change in the overall structure of the RNA. ADAR1 is known primarily edits Alu elements in RNA polymerase II (pol II) transcribed mRNAs [151]. In this study editome information for integrated Alu elements is available on the basis of multiple sequence alignment (MSA) of all Alu integrations. I used this information to perform a computational prediction of human SRP RNA structures

Predicted BSJs	4551
RNase R enriched BSJs	933
Nuclear enriched BSJs	66
RNA interaction related BSJs	16

Table 4.3: **Summary of the BSJ enrichment analysis** *from the comparison between RNAse R treated and untreated libraries from the H1ESC dataset. Supplemented by data from nuclear fractionation.*

with RNAfold [152]. However, this MSA has multiple gaps in its sequence, making RNA structure folding problematic. I manually copied all the editing sites onto the sequence of the SRP RNA. I used the editome information to compile a set of hard constrains for MFE structure prediction of the SRP RNA. I used the constrain options of RNAfold to predict two potential alternative structures with different constraints: 1) editing sites are ensured to participate in base pairing; 2) editing sites are restricted from pairing.

4.6 Results

4.6.1 BSJ enrichment analysis results

The summary of the results of the BSJ enrichment tests can be seen in Table 4.3 and Figure 4.3. There are two BSJs that are enriched in both the SPLASH data and the nuclear fraction. One of the BSJs corresponds to the *FIRRE* gene and is among the most abundant BSJs in our RNA-seq data sets. Based on the assembly with CYCLeR, there are multiple circRNA isoforms that match this BSJ and even more that originate from the same host gene. On the basis of the abundance estimation with CYCLeR, we can say that the levels of circular and linear *FIRRE* isoforms have comparable quantities. All other circRNAs enriched in the nucleus have a substantially lower ratio of the quantity of circRNA versus linear RNA and a generally lower abundance. From the circRNAs enriched in the SPLASH data, the only other example of highly abundant RNA is *CDR1as*–a circRNA with a known function as a miRNA sponge in neural cells [12]. The *CDR1as* gene is known to produce exclusively a circular isoform. CircCDR1as is the top-ranked circRNA in the SPLASH enrichment.

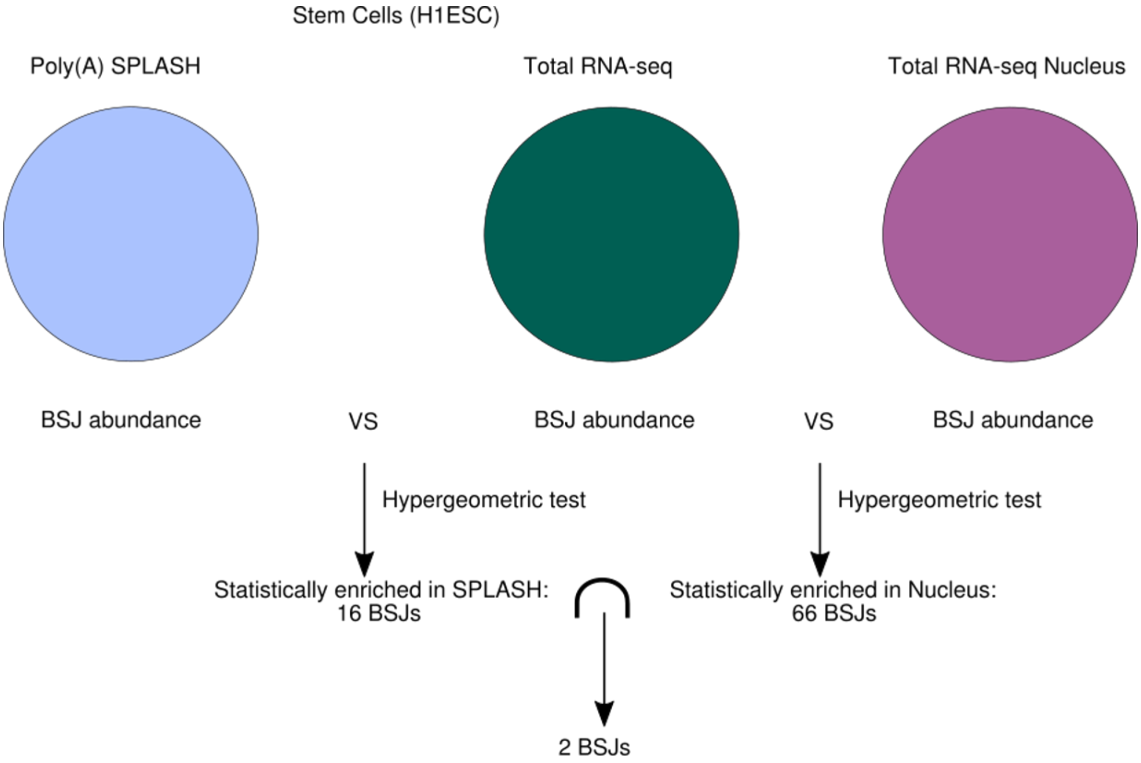


Figure 4.3: **BSJ enrichment analysis scheme for H1ESC data** *Comparison between the BSJ-spanning reads from full SPLASH library and the total ribo-depleted RNA-seq allows us to identify a set of circRNA that very likely participate in interactions with polyA-transcripts. The over-representation test gives us a list of 16 significant circRNA. A similar procedure is repeated for comparison between total ribo-depleted RNA-seq and the nuclear fraction RNA-seq. The intersection of the sets of enriched BSJ gives us circRNAs with potential functions involving RRIs in the nucleus.*

4.6.2 Clustering results

Most of the transcripts in the clusters follow two trends. They either increase in transcript levels over time after RA treatment or decrease. It is a rare occurrence for a transcript to be specific to a particular time point (only 12% of transcripts). Based on gene ontology enrichment analysis, the transcripts that correspond to an increase over time can be separated into two categories: clusters specific to a differentiation process (e.g. neural differentiation, heart muscle differentiation, keratinocyte differentiation) or immune response to outside stimuli. For most of the clusters where transcript levels drop over time, topGO cannot assign a reliable functional enrichment. However, the gene ontology analysis of two of the clusters reliably relates them to pluripotency. There are clusters specific to day 1, day 3 and HERVH-high cells. Separation of the replicates in two batches proves fruitful by facilitating the segregation of transcripts that are unique to day 1 (or days 1 and 3) of the second round of replicates in their own clusters. There are many transcripts that do not cluster due to low variance. This includes the circles of *CDR1as* and *FIRRE*, that are constitutively expressed across all samples.

The combination of co-expression clustering and RNA-RNA search provides an overabundance of results. To make the workload manageable, I focused solely on the transcripts that are co-expressed with the more abundant isoforms of the pluripotency TFs. Focusing on these specific clusters allows us to use previously available data and information focused on ESCs. As mentioned previously, we have two clusters enriched for genes related to pluripotency. One of the clusters contains all the primary (most abundant isoforms) of the typical pluripotency markers *NANOG, SOX2 and POU5F1.* The second one contains the less abundant isoforms of those genes. The two clusters could be merged and considered one due to their proximity in the original dissimilarity tree prior to cutting. Nevertheless, I decided to keep them separate and focus on the more potent cluster.

The TFs in the cluster of interest are *NANOG, SOX2, POU5F1, ZIC3, ZSCAN10, FOXH1, ETV4.* The products of the *ESRG* gene, a very specific marker of pluripotency, are also part of the cluster. The cluster also provides enrichment for endoderm differentiation. This is due to the fact that some of the endoderm markers overlap with the pluripotency markers: *DNMT3B, NANOG, SOX2, POU5F1, ZIC3, JARID2, FOXH1, CDYL.*

Interestingly, the circRNA with the highest variance in our dataset (referred to as circLARP7) also belongs to that cluster. This fact is increasingly more intriguing when we take into account that most of the transcripts produced by the *LARP7* locus, both sense and antisense, are assigned to a single cluster. This implies that circLARP7 has an additional form of control, separating its expression from the rest of the locus.

4.6.3 CircRNA Interactome analysis results

From the examined circRNAs only three cases merit further elaboration. The most abundant circRNA isoforms from the genes *CDR1as*, *FIRRE* and *LARP7.*

The three circRNAs of interest were examined for miRNA interactions by checking the H1ESC AGO2 PAR-CLIP-seq data. None of the circRNAs have reads mapping to their sequence, including circCDR1as. Although this circRNA is a potent miRNA sponge in neural cells and has high levels in ESC, it appears that circCDR1as has a different function in ESC. A further study of the chimeric reads of SPLASH showed that circCDR1as participates in multiple RRIs. Knowing that the *CDR1as* gene exclusively produces a circular isoform, we can safely assume that all chimeric reads originate from the circRNA commonly referred to as circCDR1as [12]. However, the number of reads supporting particular RRIs is small and inconsistent between replicates. Interestingly, while the levels of circCDR1as are quite consistent across our time series experiment, the number of reads supporting RRIs decreases over time, see Figure 4.4.

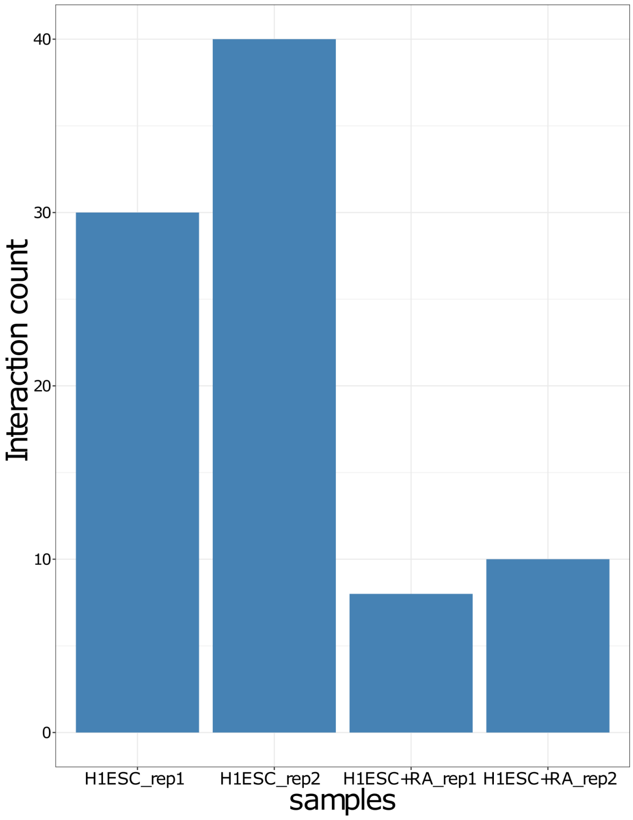


Figure 4.4: *Bar plot of the normalised number of interaction-supporting reads for circCDR1as in H1ESC SPLASH data. The number of chimeric reads supporting RRIs per library is divided by the size of the library and multiplied by a factor for convenient visualisation.*

On the basis of the BSJ enrichment analysis, only one candidate, in addition to circCDR1as, proves interesting, the most abundant circRNA isoforms of the *FIRRE* gene (referred to as circFIRRE). CircFIRRE has multiple sequences that are shown to interact with SRP RNAs (particularly *RN7SL1* and *RN7SL2*). It is important to note SRP RNAs often act as placeholders for reads that map to unannotated Alu element integrations. CircFIRRE is also enriched in the nucleus, which can be explained by the binding sequence of hnRNPU, a protein that is known to contain RNAs within the nucleus. It should be noted that these interactions would have been overlooked by a standard pipeline for chimeric read processing. The RRI sites of *FIRRE* to the specific *RN7SL1* site are multiple and vary in sequence. However, I computationally identified the consensus sequence of the RRI, see Figure 4.7.

CircLARP7 has no reads supporting RRI in the SPLASH data. A pure computational search for RRI for circLARP7 is unnecessary, as the transcript has a very long complementary sequence with the *MIR302CHG* transcripts–800 nucleotides, if the interaction occurs with the nascent *MIR302CHG* transcript. In fact, the retained introns in the circRNA sequence and the proximity of the transcription almost guarantee an interaction with the nascent *MIR302CHG* transcript.

4.6.4 Potential functional circRNAs

4.6.4.1 CircFIRRE

The *FIRRE* gene is known to have multiple integrated repetitive elements [150]. Of these repeats, there is one that contains a binding sequence for hnRNPU. This repeat is present in all *FIRRE* isoforms, often in multiple occurrences, and is logically used as a target for *FIRRE* KD. Thus, it is straightforward to assume that the *FIRRE* KD targets both linear and circular isoforms with equal success. The consistent result of *FIRRE* KD in human cells is the increase in genes related to the cellular viral response. Interestingly, KD of the *FIRRE* gene in mouse cells does not cause the same effect. If we assume that the interactions between *FIRRE* transcripts and transcripts with Alu integrations are functionally relevant, then we can explain the difference between species with the following. While in human the majority of the repetitive element integrations are attributed to the Alu family, in mouse the repetitive element integration is more diversified, even if mice do have an equivalent of the Alu family – the B1 family. Thus, any mechanism related to the processing of Alu elements would be human-specific.

The cellular virus response operates through the detection of long RNA duplexes. The distinction between foreign RNAs is facilitated by the MDA5 protein, a cytoplasmic virus response element that targets dsRNA structures [151, 153]. Human Alu-containing RNAs often form dsRNA helices. The way human RNA structures avoid the MDA5 response is by ADAR modifications of Alu-derived sequences. KD of ADAR leads to a type I interferon response similar to that of *FIRRE* KD. This leads to the hypothesis that the interaction of circFIRRE with Alu-specific sequences is related to the ADAR editing process.

According to [153], Alu-containing transcript undergo RNA editing to mimic the final structure of the SRP RNA. I used the sequence of the human SRP RNA and its known structure as part of the SRP complex as a placeholder for further analysis. According to the known SRP RNA structure, the predicted RRI sites with *FIRRE* are in fact part of a structure duplex. The only explanation is that the *FIRRE* transcripts bind to an alternative structure of the SRP RNA. Combining the fact that RNA editing changes the structure of the RNA and that the binding sites of circFIRRE overlap with duplexes of the final SRP RNA structure leads me to believe that circFIRRE serves as a chaperon, ensuring the correct folding of ALU-containing transcripts, see Figure 4.7.

As a proof of concept, I performed an editome-driven alternative folding of the SRP sequence, where the editing sites are used to restrict folding. The two alternative structures differ in the locations of stems and loops. The structure corresponding to the unedited version of the SRP RNA has open loops at the *FIRRE* RRI sites. The structure corresponding to the edited version of the SRP RNA matches the known SRP structure very well, which is a welcome surprise, given that SRP is an RNA notoriously hard to fold, due to the presence of multiple base pairs that are possible only in 3D.

What is left to determine is whether a functional RRI should be attributed to the circular or linear isoform of *FIRRE*. Both isoforms are exclusive to the nucleus as a consequence of the binding sites for hnRNPU. This protein has binding sites for the simultaneous binding to DNA and RNA, keeping RNAs anchored to the nucleus. On closer inspection, we can see that there is an exon that is present only in circular isoforms and not in any linear isoforms, see Figure 4.6. Given the low coverage of that exon, this cannot be part of the primary functional isoform, but it does hint that circular *FIRRE* isoforms have a higher RRI potential.

I also performed an equivalent analysis of *Firre* in PARIS data for mouse ESC. PARIS is an alternative RNA proximity ligation technique with a duplex selection procedure different from that of SPLASH. The difference in the experimental protocol leads to shorter chimeric fragments produced by PARIS, see Section 2.5. This meant that I cannot unambiguously map the chimeric reads originating from repetitive element integrations. Therefore, I have no way of confirming whether mouse *Firre* transcripts interact with B1 elements or not.

4.6.4.2 CircLARP7

The *LARP7* gene gives rise to multiple circRNA isoforms. Some of these isoforms have an obvious potential as microRNA sponges, due to the fact that the *LARP7* circRNA locus is on the opposite strand of the miR-302/367 cluster. This miRNA cluster is of extreme importance for pluripotency [154]. However, the regions with perfect complementarity to the microRNA sequences are not present in the most abundant *LARP7* isoform (referred to as circLARP7), see fig. 4.9. There are no detected RNA-RNA interactions in the SPLASH data for circLARP7.

An interesting feature of this isoform is high intron retention. This ensures complementarity to the nascent *MIR302CHG* transcript. The microRNAs from the miR-302/367 cluster are most likely processed as an intron of the *MIR302CHG* transcript. Furthermore, circLARP7 is enriched in the nucleus, which is a logical location of a splicing control element. circLARP7 is by a significant margin the most abundant circRNA in the pluripotency co-expression cluster. While linear *LARP7* isoforms maintain a moderate level of expression after RA treatment, circLARP7 is almost completely depleted. There are only two AGO2 CLIP-seq reads mapping to *LARP7* locus, an indication that circLARP7 is not operating as microRNA sponge.

The aforementioned results make circLARP7 a prime candidate for further experimental verification as a key element driving pluripotency.

4.7 Discussion

I employed two alternative approaches to identify functional circRNA RRIs. One is based on the co-expression and co-localisation analysis of the expression levels across transcripts in RNA-seq data. The second is multilevel examination of RNA proximity ligation data – SPLASH. Sadly, the results of functional circRNA RRIs with different methods do not have a perfect overlap. Nevertheless, both approaches identify promising candidates for further study.

The co-expression analysis supplemented by co-localisation information points to circLARP7 as the most likely participant in functional RRI. Circularisation of circLARP7 is likely partially triggered by low complexity repeats (TTATAA)n and (TTAA)n repeats in the flaking introns. However, as the circular isoform does not correlate highly with the expression of linear LARP7, there must be an additional control mechanism. This mechanism is most likely a splicing factor, whose expression is influenced by the core transcription factors driving pluripotency. The gene locus of circLARP7 has high conservation, and circRNA databases verify the presence of circLARP7 isoforms across multiple species and tissues. The location of the miR-302/367 cluster, anti-sense of *LARP7* intron, is also conserved across species. An RRI between a LARP7 transcript and the miR-302/367 cluster would justify the genes evolving with anti-sense localisation in the genome. Additionally, both genes promoter regions are binding sites for Nanog, Oct4 and Sox2, making both genes expresses under the same conditions.

The alternative approach of finding functional circRNAs based on enrichment in SPLASH data also achieved results. *FIRRE* is a conserved gene, whose unique feature is the presence of multiple integrations of repetitive elements. *FIRRE* transcripts have been proposed to act as RNA scaffolds that facilitate a higher-order DNA structure. However, KD of *FIRRE* in cell lines of different species (mouse and human) leads to different effect on the cells, puzzling researchers. My work manages to explain the difference by linking the human circFIRRE transcript to interaction with repetitive element integrations of the Alu family, which is abundant in humans as opposed to its equivalent in mice. According to the mechanism proposed, circFIRRE acts as an anchor for RNA with Alu elements and ensures that only edited RNAs can leave the nucleus, thereby preventing the trigger of the cell viral response mechanism.

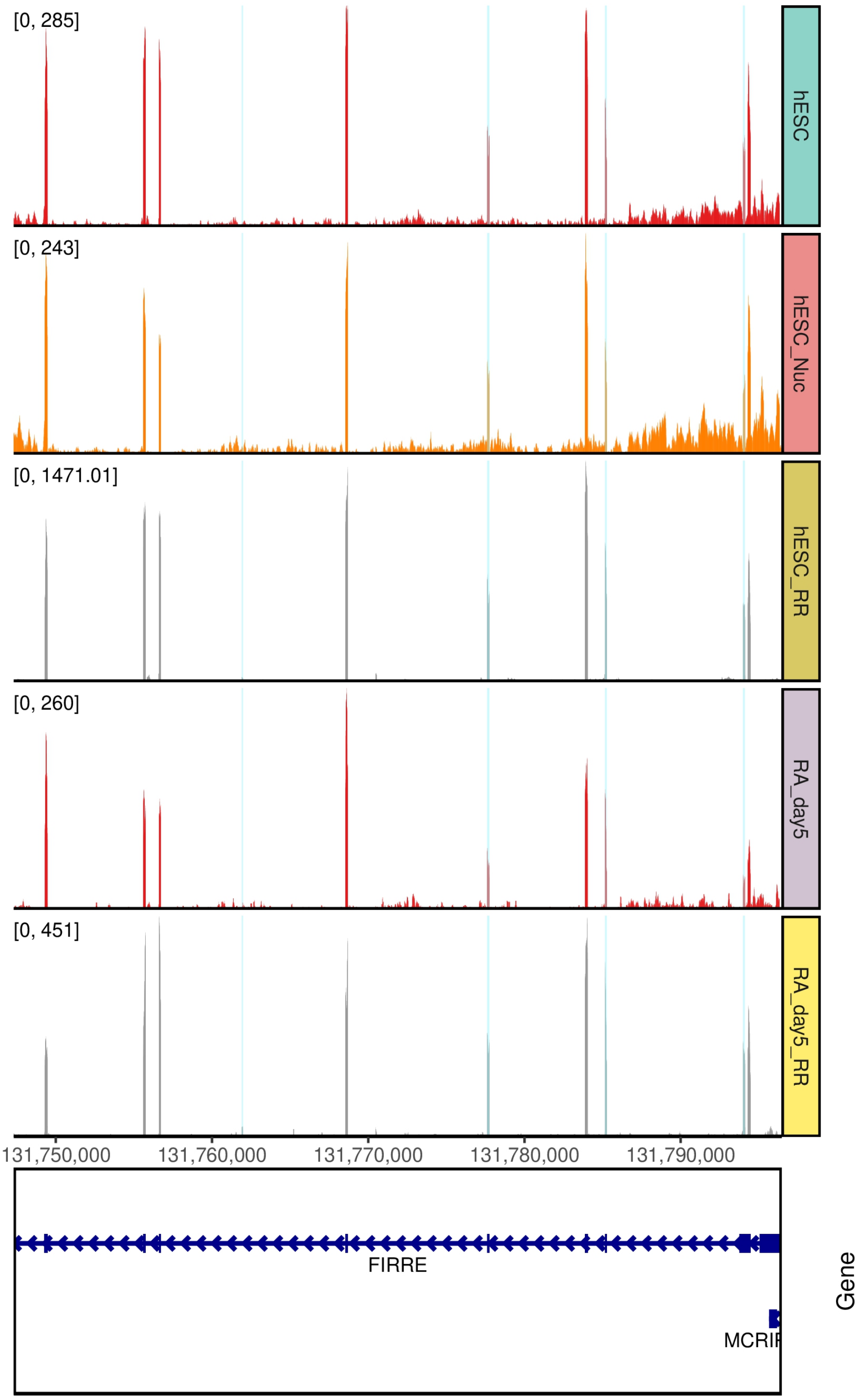


Figure 4.5: *Coverage plot of the FIRRE circRNA locus. The tracks show in order hESC, hESC nuclear fraction, hESC+RNaseR treatment, hESC on day5 of RA treatment, hESC on day5 of RA treatment+RNaseR treatment. The plot shows only the locus of the circRNA. Both linear and circular FIRRE isoforms are exclusive to the nucleus. The interaction regions of the ALU-specific sequences are marked with light blue. The regions are based on the chimeric read sequence. One of the interaction regions matches an exon with a lower abundance. What is interesting about this exon is the fact that it appears to be present only in circRNAs as can be seen in Figure 4.6.*

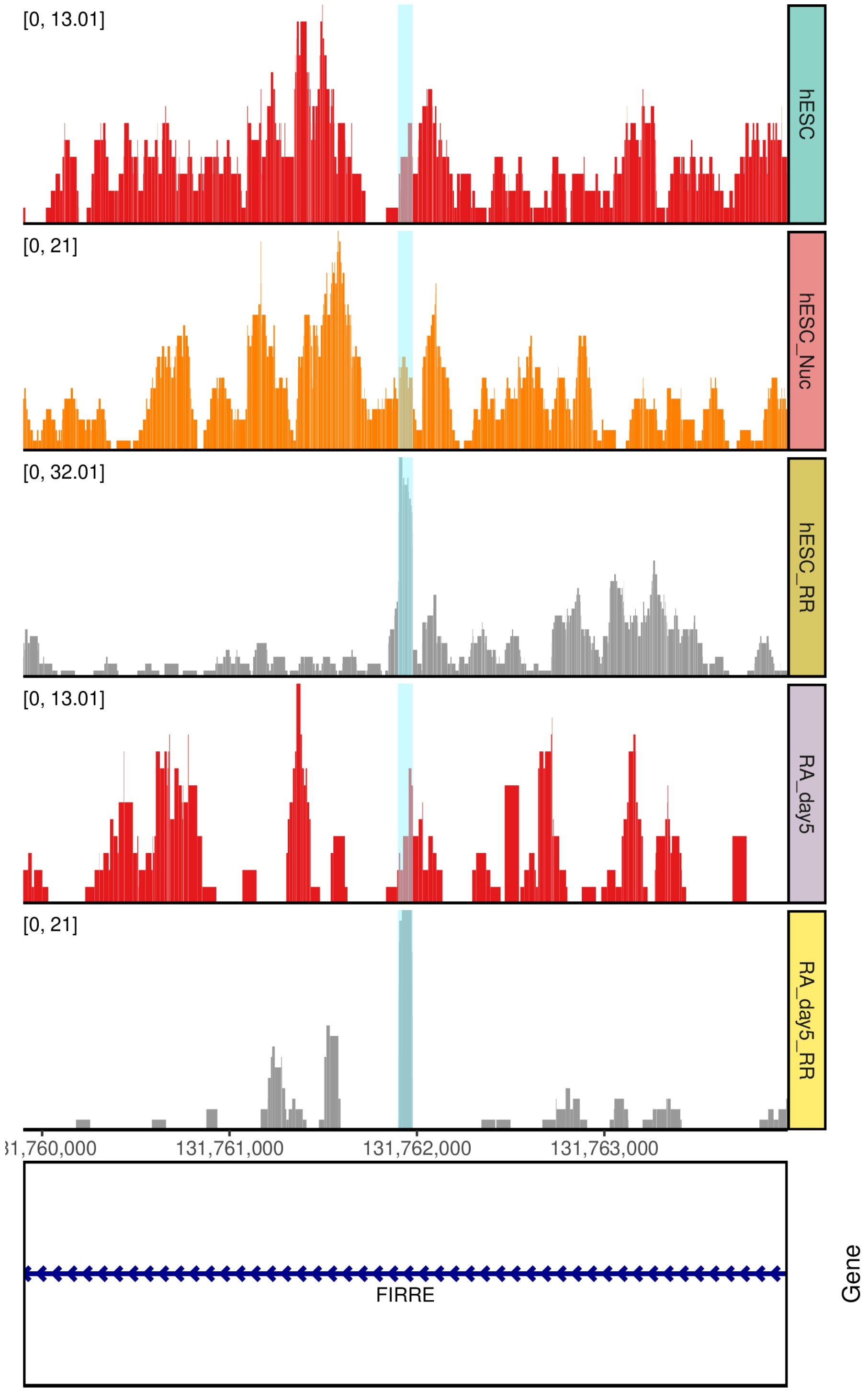


Figure 4.6: *A coverage plot showing a detailed zoom of a specific FIRRE circRNA exon. The tracks show in order hESC, hESC nuclear fraction, hESC+RNaseR treatment, hESC on day5 of RA treatment, hESC on day5 of RA treatment+RNaseR treatment. We can see the ALU interaction sequence (indicated by light blue) belonging to an exon specifically seen only in the RNase R treated library.*

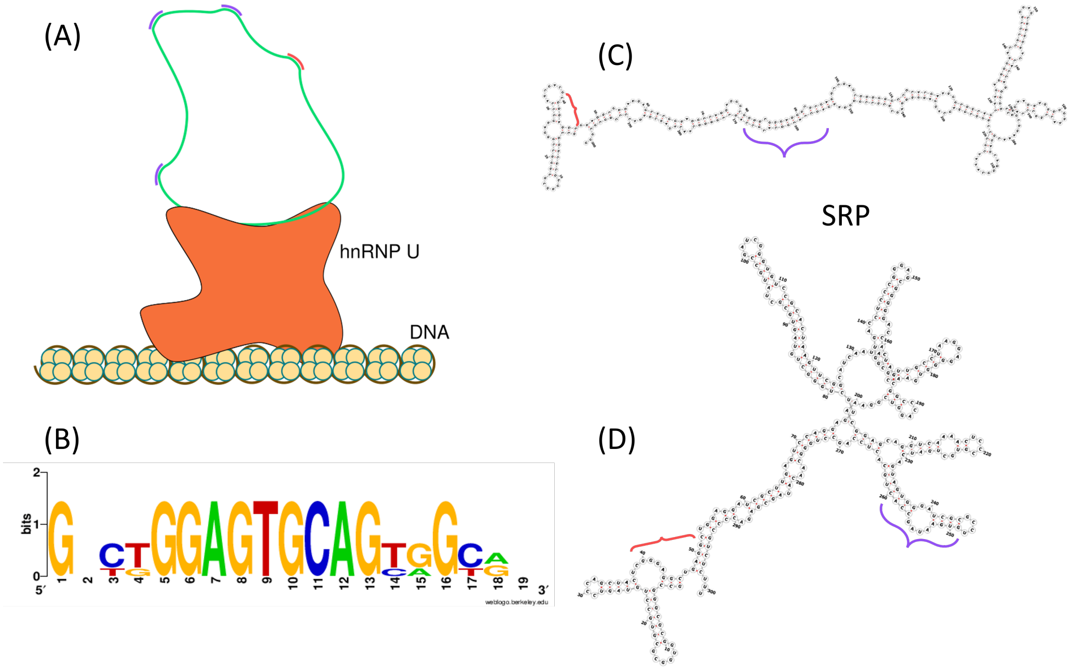


Figure 4.7: **CircFIRRE and SRP RNA binding.** *On (A) we see the interaction sites between the SRP RNA and circFIRRE. On (B) we see the sequence logo showing the consensus sequence of the multiple interaction sites in circFIRRE marked with blue. On (C) and (D) I show potential SRP alternative structures post- and pre- RNA editing. In (C) we can see the both RRI regions are inaccessible due to full participation of the sequence in duplexes. The open loops in the pre-editing folding (D) can accommodate circFIRRE binding.*

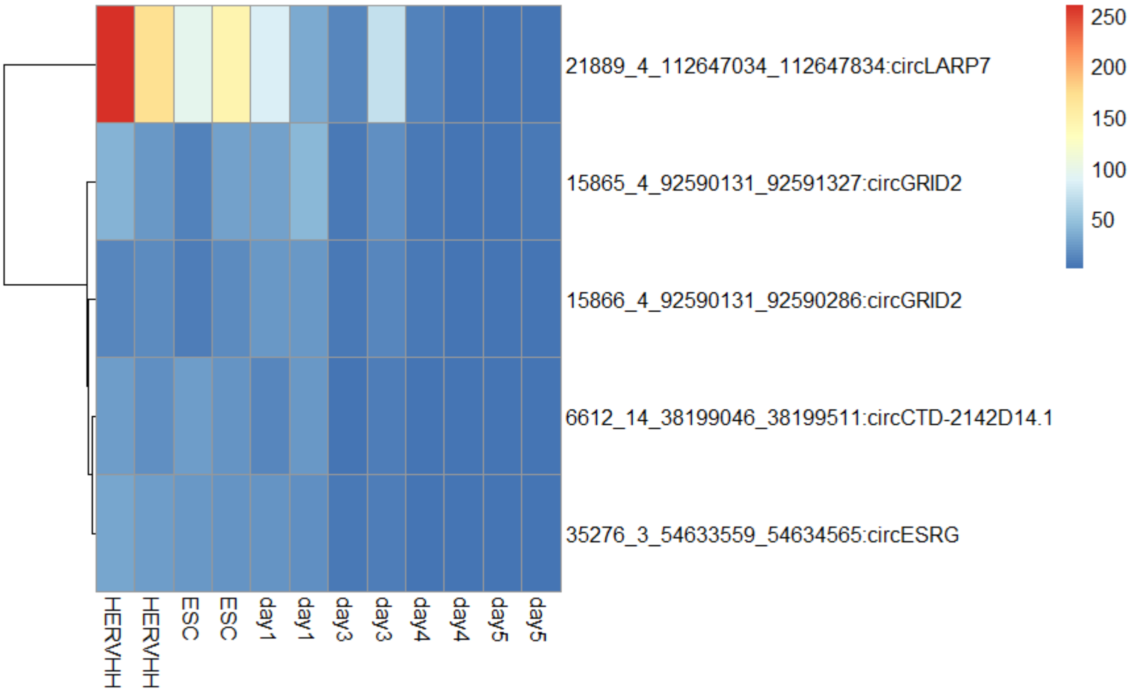


Figure 4.8: *Heatmap transcript abundance (in TPM) of the top 5 most expressed circRNA transcripts that co-express with pluripotency markers. HERVHhigh refers to HERVH-high cells. The circRNAs are represented by their unique ID, coordinates of the BSJ and host gene.*

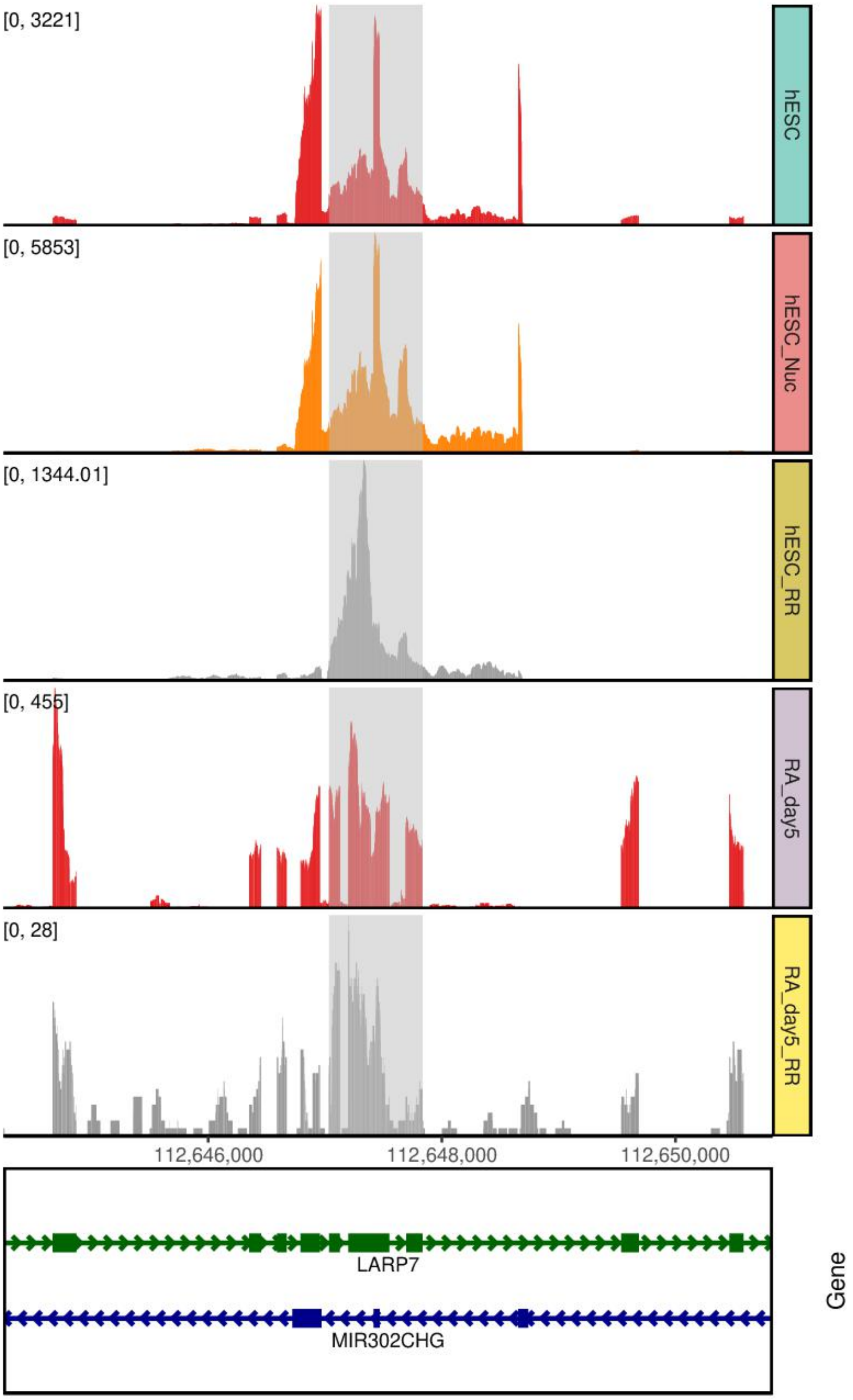


Figure 4.9: *Coverage plot of the LARP7 circRNA locus. For convenience, only single isoforms for LARP7 and MIR302CHG are shown. The tracks show in order hESC, hESC nuclear fraction, hESC+RNaseR treatment, hESC on day5 of RA treatment, hESC on day5 of RA treatment+RNaseR treatment. The primary circRNA locus is marked with gray. Note: the noticeably high peaks in the hESC tracks correspond to circLARP7 and MIR302CHG. Interestingly, linear LARP7 transcripts have no noticeable presence in the nuclear fraction. Additionally, on day 5 after RA treatment, there are neither circLARP7 nor MIR302CHG transcripts present. There are mutiple BSJ attributed to the LARP7 gene. On the figure, we mostly see 2 circRNA contributing to the coverage: our circRNA of interest marked with gray band and one monoexonic circRNA.*

Chapter 5

Results and Discussion overview

5.1 Results summary

The first step towards achieving my goal of identifying functional circRNA-RNA interactions was to design a tool for identification of the full sequence of a circRNA. CYCLeR succeeds in reconstruction of circRNA transcript, in cases where alternative tools fail. CYCLeR manages to identify circular isoforms with very low false-positive rate, without a limit on the transcript length. CYCLeR also succeeds in the assembly of circRNAs with unannotated transcript features. An additional advantage of the pipeline I developed is the simultaneous quantification of linear and circular RNAs. I have shown that the improved quantification strategy of CYCLeR leads to a higher chance of identifying the functional association of a circRNA.

Enrichment of circRNA participating in interactions supported by SPLASH data pointed to circFIRRE as an interesting candidate. circFIRRE is a constitutively expressed circRNA that is considered the predominant isoform of the *FIRRE* gene. All isoforms of the *FIRRE* gene are exclusively localised in the nucleus, due to repeat sequence that binds to the hnRPU protein. The duplexes identified from the SPLASH data show multiple interactions between the *FIRRE* transcripts and Alu-containing transcripts. KD of the transcripts of the gene pushes the human cells to go into a viral response state. However, the effect of KD of *Firre* in mouse is completely different, indicating that *FIRRE* has a human specific function. The fact that Alu elements are occurring in human more often than their equivalent in mouse can explain the difference in response to KD between the two organisms. Alu elements are known to trigger viral response, when the editing mechanism of the cell is perturbed. Furthermore, structural analysis using the SRP RNA as a placeholder for integrated Alu sequences shows that the interaction between circFIRRE and Alu element can occur only in unedited sequence. Therefore, the viral response in humans upon *FIRRE* KD can be attributed to *FIRRE* being an integral element of RNA editing process, ensuring that Alu-containing transcripts do not leave the nucleus without being edited. We generate a time-series data set to identify circRNAs, which co-express with known markers of pluripotency. Based on the membership in the cluster with transcripts known to affect pluripotency, I managed to determine the most likely candidate for circRNA that has effect on pluripotency. The nuclear fraction RNA-seq results allow us to pinpoint the localisation of this circRNA and suggest its potential function. The high level of intron retention in circRNA gives it a unique RRI potential that cannot be replicated by the linear isoforms of *LARP7*. The anti-sense localisation of the *LARP7* and *MIR302CHG* genes and their simultaneous expression ensures at least temporary co-localisation of the transcripts originating from the two genes. Additionally, the fact that linear LARP7 is exported from the nucleus, while *MIR302CHG* transcripts and circLARP7 remain in the nucleus is an additional hint for a functional interaction. The sequences of the retained introns of circLARP7 overlap the intronic regions of the *MIR302CHG* gene; therefore, an interaction would occur between circLARP7 and the nascent *MIR302CHG* transcript. My hypothesis is that circLARP7 mediates the processing of the nascent *MIR302CHG* transcript and facilitates the production of miRNAs from the miR-302/367 cluster.

5.2 Discussion

To identify circRNA-specific RRIs, it is essential to know the full sequence of all isoforms that correspond to a BSJ. The accent of the isoform identification should be on the identification of exons or retained introns that are present in circRNAs while absent linear RNAs. The full sequence of the circRNA transcripts is also necessary for transcript quantification and downstream processing via differential expression or co-expression analysis. CYCLeR provides an efficient and robust solution to both of those objectives. However, the tool has downsides. The need for circRNA-enriched RNA-seq libraries increases the work load and cost of an experiment. Nevertheless, I have shown that those the sample number can be minimised to only a few key time points of the study, and the results will still be informative.

The fact that circRNAs are assembled separately from linear leaves room for errors in the linear transcript assembly. While CYCLeR performs a step for selection of features that are specific for circRNAs, linear RNA assembly tools do not have such a step. Therefore, reads corresponding to circRNA can influence the assembly of linear RNA. To avoid such issues, I have used only known linear annotation. As my studies have focus only on commonly used cell lines, the likelihood of the presence of highly relevant unknown linear isoforms is very low. Furthermore, there is an added advantage to keeping the assembly of linear and circular transcript separate. Due to the higher relative expression levels of linear RNA, many circRNA would fall under the standard detection limits that navigate robust linear transcript assembly. A hybrid method that uses both short-read and long-read data cannot yet be developed for circRNA, because the output of protocols for long-read circRNA data is not reliable. Thus, CYCLeR is the best possible option for circRNA transcript assembly.

The abundance estimation by EM that CYCLeR employs is most likely the optimal approach toward simultaneous quantification of linear and circular RNA. The tools that utilise EM abundance estimation have the same core algorithm. The difference in output can be attributed to different approaches towards RNA-seq bias correction. Even if a novel improved tool for transcript quantification is developed, CYCLeR can easily be adjusted to work with any novel tool, because the final step of the quantification is not carried out by the core CYCLeR package, but performed separately.

The purpose of the development of CYCLeR was to facilitate my research on the RRI interaction that involves circRNA. Extensive benchmarks have proven that CYCLeR is the optimal tool for circRNA assembly. Furthermore, the quantification provided by CYCLeR in combination with steps from the WGCNA pipeline allowed identification of circRNA-specific patterns in a fruit fly dataset. When the same strategy was applied to the data generated specifically for this study, the results exceeded expectations. All major markers of pluripotency were assigned to a single co-expression cluster, making the search for circRNA with effect on pluripotency a straightforward task. This result is in part due to the data generation and, particularly, the consistent and low fragment size of the RNA-seq libraries. The only downside in the experimental design is the localisation data. To minimise spending, the experimental design contains only RNA-seq data for the nuclear fraction and no matching library from the cytosolic fraction. Although comparison with the whole cell library allows identification of circRNAs enriched in the nucleus, we do not know if their localisation is exclusive to the nucleus. However, exclusive nuclear localisation is not necessary for participation in RRIs, therefore, this oversight is not detrimental to the study.

As a candidate circRNA with a major effect on pluripotency, I have selected circLARP7. My primary hypothesis is that circLARP7 affects the processing of *MIR302CHG* miRNA cluster. The combination of these anti-sense genes is very well conserved, leading to similar AS patterns across species. Future experiments are needed to verify the hypothesis. We are already aware that the transcripts are co-localised at least transiently; therefore, fluorescent in situ hybridisation will give very little new information. KD of the circRNA is the only option to test the hypothesis. Usually, the target for circRNA KD is mainly the BSJ. However, circLARP7 has circRNA-specific intron retention that provides more target sites for KD. Alternatively, the interaction partners of circLARP7 can identified with a dedicated pull-down experiment.

CircFIRRE is a constitutively expressed circRNA, which indicates that its potential function is a common necessity in all cells. The fact that the RRI of interest is between two transcripts related to repetitive elements makes any experimental verification difficult. Attempting to visualise transcripts with relation to the Alu family would net an over-saturated signal. In previous studies of the *FIRRE* gene, a repeat that is present multiple times at the *FIRRE* locus is used as a target. However, this repeat is present in both the linear and circular isoforms. The publicly available data are poly(A)-enriched, hence lacking information about the state of the circRNA levels. The sequence around the BSJ is not unique enough to serve as a target site for a KD experiment. A possible strategy could be to target the sequences present in both linear and circular isoforms and select a probe that preferentially affects the circRNA.

The editome-based folding described in Section 4.5.2, is a proof-of-concept test to test the use of editing sites to predict alternative structure of the SRP molecule. While the test manages to provide insight into the mechanism of action of circFIRRE the analysis needs to be adapted to dedicated data. The future analysis needs to be performed on samples with KD of circFIRRE versus control samples. All integrated Alu sequences should be juxtaposed into a MSA with additional information on RNA editing sites. Then editome-based folding needs to be performed with algorithm designed for MFE folding based on evolutionary information. Follow-up experiments and analytical procedures have the potential to reveal a novel function of circRNA and more importantly a novel type of cellular mechanism.

The main results of my doctoral work are the development of CYCLeR and using it for identification of novel circRNA-RNA interactions. Both CYCLeR and the co-expression analysis of time-series data of RA treatment of ESCs worked exactly as predicted. CYCLeR is made available as a publicly accessible Docker image that is well documented and easy to use. The results of the co-expression clustering are not limited to information about pluripotency. The efficient clustering has the unexplored potential to provide insight into pathways related to differentiation. Thus, both the method development part of the project and the data analysis results of the project can be used in future scientific studies.

Chapter 6

Publications accepted during doctoral work

Stefan R. Stefanov, Irmtraud M. Meyer (2018) Deciphering the Universe of RNA Structures and trans RNA–RNA Interactions of Transcriptomes In Vivo: *From Experimental Protocols to Computational Analyses. In: Rajewsky N., Jurga S., Barciszewski J. (eds) Systems Biology. RNA Technologies. Springer, Cham* DOI: 10.1007/978-3-319-92967-5_9

Peter Menzel, Alexandra L. McCorkindale, **Stefan R. Stefanov**, Robert P. Zinzen , Irmtraud M. Meyer (2019): *Transcriptional dynamics of microRNAs and their targets during Drosophila neurogenesis*, *RNA Biology*, DOI: 10.1080/15476286.2018.155890

Masin Abo-Rady, Norman Kalmbach, Arun Pal, Carina Schludi,..., **Stefan R. Stefanov**,..., Jared L.Sterneckert (2020): *Knocking out C9ORF72 Exacerbates Axonal Trafficking Defects Associated with Hexanucleotide Repeat Expansion and Reduces Levels of Heat Shock Proteins*, *Stem Cell Reports*, DOI: 10.1016/j.stemcr.2020.01.010

Stefan R. Stefanov, Irmtraud M. Meyer *CYCLER–a novel tool for the full isoform assembly and quantification of circRNAs Nucleic Acids Research*, DOI: 10.1093/nar/gkac1100

Bibliography

[1] Black DL (2003) Mechanisms of alternative pre-messenger rna splicing. Annual Review of Biochemistry 72(1):291–336

[2] Gehring NH, Roignant JY (2021) Anything but ordinary – emerging splicing mechanisms in eukaryotic gene regulation. Trends in Genetics 37:355–372

[3] Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: Towards a cellular code. Nature Reviews Molecular Cell Biology 6:386–398

[4] Fica SM, Tuttle N, Novak T et al (2013) Rna catalyses nuclear pre-mrna splicing. Nature 503:229–234

[5] Shi Y (2017) Mechanistic insights into precursor messenger rna splicing by the spliceosome. Nature Reviews Molecular Cell Biology 18:655–670

[6] Trotta CR, Miao F (1997) The yeast trna splicing endonuclease: A tetrameric enzyme with two active site subunits homologous to the archaeal trna endonucleases. Cell 89:849–858

[7] Ge J, Liu H, Yu YT (2010) Regulation of pre-mrna splicing in xenopus oocytes by targeted 2'-o-methylation. RNA 16:1078–1085

[8] Tilgner H, Knowles DG, Johnson R et al (2012) Deep sequencing of subcellular rna fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncrnas. Genome Research 22:1616–1625

[9] Ashwal-fluss R, Meyer M, Pamudurti NR et al (2014) Article circrna biogenesis competes with pre-mrna splicing. MOLCEL 56:55–66

[10] Ullah F, Hamilton M, Reddy AS, Ben-Hur A (2018) Exploring the relationship between intron retention and chromatin accessibility in plants. BMC genomics 19:21

[11] Salzman J, Gawad C, Wang PL et al (2012) Circular rnas are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS ONE 7:e30733

[12] Memczak S, Jens M, Elefsinioti A et al (2013) Circular rnas are a large class of animal rnas with regulatory potency. Nature 495:333–338

[13] Szabo L, Salzman J (2016) Detecting circular rnas : bioinformatic and experimental challenges. Nature Reviews Genetics 17:679–692

[14] Cape B, Swain A, Nicolis S et al (1993) Circular transcripts of the testis-determining gene sry in adult mouse testis. Cell 73:1019–1030

[15] Zhang Y, ou Zhang X, Chen T et al (2013) Article circular intronic long noncoding rnas. Molecular Cell 51:792–806

[16] Suzuki H, Zuo Y, Wang J et al (2006) Characterization of rnase r-digested cellular rna source that consists of lariat and circular rnas from pre-mrna splicing. Nucleic Acids Research 34

[17] Li Z, Huang C, Bao C et al (2016) Exon-intron circular rnas regulate transcription in the nucleus. NATURE STRUCTURAL & MOLECULAR BIOLOGY 22:256–264

[18] Yang Y, Fan X, Mao M et al (2017) Extensive translation of circular rnas driven by n6 -methyladenosine. Nature Publishing Group 27:626–641

[19] F S, S N, AR. C (1977) Dna sequencing with chain-terminating inhibitors. PNAS 12:5463–5467

[20] Smith LM, Fung S, Hunkapiller MW et al (1985) The synthesis of oligonucleotides containing an aliphatic aio group at the 5' terminus: synthesis of fluorescent dna primers for use in dna sequence analysis. Nucleic Acids Research 13:2399–2412

[21] Wang Z, Gerstein M, Snyder M (2009) Rna-seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics 10:57–63

[22] Chu Y, Corey DR (2012) Rna sequencing: Platform selection, experimental design, and data interpretation. Nucleic Acid Therapeutics 22:271–274

[23] Author M, Schena M, Shalon D et al (1995) Quantitative monitoring of gene expression patterns with a complementary dna. Science 270:467–470

[24] Shalon D, Smith SJ, Brown PO (1996) A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. Genome Research 6:639–645

[25] Rao MS, Vleet TRV, Ciurlionis R et al (2019) Comparison of rna-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. Frontiers in Genetics 10

[26] Deng ZL, Münch PC, Mreches R, McHardy AC (2022) Rapid and accurate identification of ribosomal rna sequences via deep learning. Nucleic Acids Research 50:E60

[27] Griffith M, Walker JR, Spies NC et al (2015) Informatics for rna sequencing: A web resource for analysis on the cloud. PLoS Computational Biology 11

[28] Jeck WR, Sorrentino JA, Wang KAI et al (2013) Circular rnas are abundant , conserved , and associated with alu repeats. RNA pages 141–157

[29] Hrdlickova R, Toloue M, Tian B (2017) Rna-seq methods for transcriptome analysis. Wiley Interdisciplinary Reviews: RNA 8(1):e1364

[30] Metge F, Czaja-hasse LF, Reinhardt R, Dieterich C (2017) Fuchs — towards full circular rna characterization using rnaseq. PeerJ pages 1–14

[31] Illumina workflow.

[32] Conesa A, Madrigal P, Tarazona S et al (2016) A survey of best practices for rna-seq data analysis. Genome Biology 17

[33] Kim D, Pertea G, Trapnell C et al (2013) Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14:R36

[34] Dobin A, Davis CA, Schlesinger F et al (2013) Star: Ultrafast universal rna-seq aligner. Bioinformatics 29:15–21

[35] Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv 00:1–3

[36] Kim D, Salzberg SL (2011) Tophat-fusion: an algorithm for discovery of novel fusion transcripts. Genome biology 12:R72

[37] Westholm JO, Miura P, Graveley BR et al (2014) Genome-wide analysis of drosophila circular rnas reveals their structural and sequence properties and age-dependent neural accumulation. CellReports 9:1966–1980

[38] Gao Y, Wang J, Zhao F (2015) Ciri : an efficient and unbiased algorithm for de novo circular rna identification. Genome Biology pages 1–16

[39] Eid J, Fehr A, Gray J et al (2009) Real-time dna sequencing from single polymerase molecules. Science 323:133–138

[40] Levene MJ, Korlach J, Turner SW et al (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299:682–686

[41] Niedringhaus TP, Milanova D, Kerby MB et al (2011) Landscape of next-generation sequencing technologies. Analytical Chemistry 83:4327–4341

[42] You X, Vlatkovic I, Babic A et al (2015) Neural circular rnas are derived from synaptic genes and regulated by development and plasticity. Nature Publishing Group 18:603–610

[43] Zhang J, Hou L, Zuo Z et al (2021) Comprehensive profiling of circular rnas with nanopore sequencing and ciri-long. Nature Biotechnology 39:836–845

[44] Xin R, Gao Y, Gao Y et al (2021) isocirc catalogs full-length circular rna isoforms in human transcriptomes. Nature Communications 12:1–11

[45] Rahimi K, Venø MT, Dupont DM, Kjems J (2021) Nanopore sequencing of brain-derived full-length circrnas reveals circrna-specific exon usage, intron retention and microexons. Nature Communications 12

[46] Anders S, a Reyes , Huber W (2012) Detecting diferential usage of exons from rna-seq data. Genome Res 22:2008–2017

[47] Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28:511–5

[48] Pertea M, Pertea GM, Antonescu CM et al (2015) Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. Nature Biotechnology 33:290–295

[49] Zheng Y, Ji P, Chen S et al (2019) Reconstruction of full-length circular rnas enables isoform-level quantification. Genome Medicine 11:1–20

[50] Zhang Y, Xue W, Li X et al (2016) The biogenesis of nascent circular rnas. Cell Reports 15:611–624

[51] Wu J, Li Y, Wang C et al (2019) Circast: Full-length assembly and quantification of alternatively spliced isoforms in circular rnas. Genomics, Proteomics and Bioinformatics 17:522–534

[52] Stefanov SR, Meyer IM (2023) Cycler —a novel tool for the full isoform assembly and quantification of circrnas. Nucleic Acids Research 51:e10

[53] Patro R, Mount SM, Kingsford C (2013) Sailfish: Alignment-free isoform quantification from rna-seq reads using lightweight algorithms. Nature Biotechnology 32:462–464

[54] Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic rna-seq quantification. Nature Biotechnology 34:525–527

[55] Patro R, Duggal G, Love MI et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14:417–419

[56] Li B, Dewey CN (2011) Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinformatics 12:323

[57] Zhao Y, Li MC, Konaté MM et al (2021) Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. Journal of Translational Medicine 19

[58] Pimentel H, Bray NL, Puente S et al (2017) Differential analysis of rna-seq incorporating quantification uncertainty. Nature Methods 14:687–690

[59] Trapnell C, Hendrickson DG, Sauvageau M et al (2013) Differential analysis of gene regulation at transcript resolution with rna-seq. Nature biotechnology 31:46–53

[60] Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology 15:1–21

[61] Li J, Tibshirani R (2013) Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. Statistical Methods in Medical Research 22:519–536

[62] Robinson MD, McCarthy DJ, Smyth GK (2009) edger: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140

[63] Nowicka M, Robinson MD (2016) Drimseq: a dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Research 5:1356

[64] Langfelder P, Horvath S (2008) Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics 9:559

[65] Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: Current approaches and outstanding challenges. PLoS Computational Biology 8

[66] Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102:15545–15550

[67] O'Brien J, Hayder H, Zayed Y, Peng C (2018) Overview of microrna biogenesis, mechanisms of actions, and circulation. Frontiers in Endocrinology 9:1–12

[68] Argaman L, Altuvia S (2000) fhla repression by oxys rna: Kissing complex formation at two sites results in a stable antisense-target rna complex. Journal of Molecular Biology 300:1101–1112

[69] Salim NN, Feig AL (2010) An upstream hfq binding site in the fhla mrna leader region facilitates the oxys-fhla interaction. PLoS ONE 5:1–11

[70] Hartswood E, Brodie J, Vendra G et al (2012) Rna:rna interaction can enhance rna localization in drosophila oocytes. Rna 18:729–737

[71] Kiss T (2002) Small nucleolar rnas: An abundant group of noncoding rnas with diverse cellular functions. Cell 109:145–148

[72] Kufel J, Grzechnik P (2019) Small nucleolar rnas tell a different tale. Trends in Genetics 35:104–117

[73] Korostelev A, Trakhanov S, Laurberg M, Noller HF (2006) Crystal structure of a 70s ribosome-trna complex reveals functional interactions and rearrangements. Cell 126:1065–1077

[74] Lee FC, Ule J (2018) Advances in clip technologies for studies of protein-rna interactions. Molecular Cell 69:354–369

[75] Lunde BM, Moore C, Varani G (2007) Rna-binding proteins: Modular design for efficient function. Nature Reviews Molecular Cell Biology 8:479–490

[76] Solé A, Mencia N, Villalobos X et al (2013) Validation of mirna-mrna interactions by electrophoretic mobility shift assays. BMC Research Notes 6:2–8

[77] Palau W, Masante C, Ventura M, Primo CD (2013) Direct evidence for rna-rna interactions at the 3' end of the hepatitis c virus genome using surface plasmon resonance. Rna 19:982–991

[78] Narberhaus F (2010) Translational control of bacterial heat shock and virulence genes by temperature-sensing mrnas. RNA Biology 7

[79] Mandal M, Breaker RR (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. Nature Structural and Molecular Biology 11:29–35

[80] Nechooshtan G, Elgrably-weiss M, Sheaffer A et al (2009) A ph-responsive riboregulator. Genes & Development 23:2650–2662

[81] Pedersen JS, Meyer IM, Forsberg R et al (2004) A comparative method for finding and folding rna secondary structures within protein-coding regions. Nucleic Acids Research 32:4925–4936

[82] Pedersen JS, Forsberg R, Meyer IM, Hein J (2004) An evolutionary model for protein-coding regions with conserved rna structure. Molecular Biology and Evolution 21:1913–1922

[83] Lai D, Proctor JR, Meyer IM (2013) On the importance of cotranscriptional rna structure formation. Rna 19:1461–1473

[84] Flores JK, Ataide SF (2018) Structural changes of rna in complex with proteins in the srp. Frontiers in Molecular Biosciences 5:1–8

[85] Caines R, Cochrane A, Kelaini S et al (2019) The rna-binding protein qki controls alternative splicing in vascular cells, producing an effective model for therapy. Journal of cell science 132

[86] Mazloomian A, Meyer IM (2015) Genome-wide identification and characterization of tissue-specific rna editing events in d. melanogaster and their potential role in regulating alternative splicing. RNA Biology 12:1391–1401

[87] Saldi T, Fong N, Bentley DL (2018) Transcription elongation rate affects nascent histone pre-mrna folding and 3' end processing. Genes and Development 32:297–308

[88] Villamizar O, Chambers CB, Riberdy JM et al (2016) Long noncoding rna saf and splicing factor 45 increase soluble fas and resistance to apoptosis. Oncotarget 7:13810–13826

[89] Jin Y, Chen Z, Liu X, Zhou X (2013) Evaluating the microrna targeting sites by luciferase reporter gene assay. Methods in Molecular Biology 936:117–127

[90] Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. Journal of Molecular Biology 288:911–940

[91] Lorenz R, Bernhart SH, zu Siederdissen CH et al (2011) Viennarna package 2.0. Algorithms for Molecular Biology 6:1–14

[92] Wiebe NJ, Meyer IM (2010) Transat-a method for detecting the conserved helices of functional rna structures, including transient, pseudo-knotted and alternative structures. PLoS Computational Biology 6:1–13

[93] Rivas E (2020) Rna structure prediction using positive and negative evolutionary information. PLoS Computational Biology 16:1–25

[94] Mattei E, Ausiello G, Ferrè F, Helmer-Citterich M (2014) A novel approach to represent and compare rna secondary structures. Nucleic Acids Research 42:6146–6157

[95] Lai D, Meyer IM (2015) A comprehensive comparison of general rna-rna interaction prediction methods. Nucleic Acids Research 44:1–13

[96] Lalaouna D, Prévost K, Eyraud A, Massé E (2017) Identification of unknown rna partners using maps. Methods 117:28–34

[97] Lalaouna D, Desgranges E, Caldelari I, Marzi S *MS2-Affinity Purification Coupled With RNA Sequencing Approach in the Human Pathogen Staphylococcus aureus* volume 612 Elsevier Inc. 1 edition (2018)

[98] Kretz M, Siprashvili Z, Chu C et al (2013) Control of somatic tissue differentiation by the long non-coding rna tincr. Nature 493:231–235

[99] Engreitz JM, Sirokman K, McDonel P et al (2014) Rna-rna interactions enable specific targeting of noncoding rnas to nascent pre-mrnas and chromatin sites. Cell 159:188–199

[100] Zhang X, Shen B, Cui Y (2019) Ago hits-clip expands microrna-mrna interactions in nucleus and cytoplasm of gastric cancer

cells 06 biological sciences 0601 biochemistry and cell biology 06 biological sciences 0604 genetics. BMC Cancer 19:1–9

[101] Melamed S, Peer A, Faigenbaum-Romm R et al (2016) Global mapping of small rna-target interactions in bacteria. Molecular Cell 63:884–897

[102] Helwak A, Tollervey D (2014) Mapping the mirna interactome by cross-linking ligation and sequencing of hybrids (clash). Nature Protocols 9:711–728

[103] Nguyen TC, Cao X, Yu P et al (2016) Mapping rna-rna interactome and rna structure in vivo by mario. Nature Communications 7:1–12

[104] Cai Z, Cao C, Ji L et al (2020) Ric-seq for global in situ profiling of rna–rna spatial interactions. Nature 582:432–437

[105] Calvet JP, Pederson T (1979) Heterogeneous nuclear rna double-stranded regions probed in living hela cells by crosslinking with the psoralen derivative aminomethyltrioxsalen* (secondary structure of inverted repeat heterogeneous nuclear rna sequences/4'-aminomethyl4,5',8-trimethylpso. Biochemistry 76:755–759

[106] Lu Z, Zhang QC, Lee B et al (2016) Rna duplex map in living cells reveals higher-order transcriptome structure. Cell 165:1267–1279

[107] Aw JGA, Shen Y, Wilm A et al (2016) In vivo mapping of eukaryotic rna interactomes reveals principles of higher-order organization and regulation. Molecular Cell 62:603–617

[108] Sharma E, Sterne-Weiler T, O'Hanlon D, Blencowe BJ (2016) Global mapping of human rna-rna interactions. Molecular Cell 62:618–626

[109] Ziv O, Gabryelska MM, Lun AT et al (2018) Comrades determines in vivo rna structures and interactions. Nature Methods 15:785–788

[110] Quinlan AR, Hall IM (2010) Bedtools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842

[111] Lawrence M, Huber W, Pagès H et al (2013) Software for computing and annotating genomic ranges. PLoS Computational Biology 9:1–10

[112] Kato Y, Sato K, Hamada M et al (2011) Ractip: Fast and accurate prediction of rna-rna interaction using integer programming. Bioinformatics 27:i460–i466

[113] ou Zhang X, bin Wang H, Zhang Y et al (2014) Complementary sequence-mediated exon circularization. Cell 159:134–147

[114] Szabo L, Morey R, Palpant NJ et al (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular rna during human fetal development. Genome Biology pages 1–26

[115] Gao Y, Wang J, Zheng Y et al (2016) Comprehensive identification of internal structure and alternative splicing events in circular rnas. Nature Communications 7:1–13

[116] Li M, Xie X, Zhou J et al (2017) Quantifying circular rna expression from rna-seq data using model-based framework. Bioinformatics (Oxford, England) pages 1–9

[117] Gao Y, Zhang J, Zhao F (2018) Circular rna identification based on multiple seed matching. Briefings in bioinformatics 19:803–810

[118] Ma XK, Wang MR, Liu CX et al (2019) Circexplorer3: A clear pipeline for direct comparison of circular and linear rna expression. Genomics, Proteomics and Bioinformatics 17:511–521

[119] Zhang J, Chen S, Yang J, Zhao F (2020) Accurate quantification of circular rnas identifies extensive circular isoform switching events. Nature Communications 11

[120] Goldstein LD, Cao Y, Pau G et al (2016) Prediction and quantification of splice events from rna-seq data. PLoS ONE 11:1–18

[121] Frazee AC, Jaffe AE, Langmead B, Leek JT (2015) Polyester: Simulating rna-seq datasets with differential transcript expression. Bioinformatics 31:2778–2784

[122] Griebel T, Zacher B, Ribeca P et al (2012) Modelling and simulating article rna-seq experiments with the flux simulator. Nucleic Acids Research 40:10073–10083

[123] Hansen TB, Venø MT, Damgaard CK, Kjems J (2016) Comparison of circular rna prediction tools. Nucleic Acids Research 44

[124] Danecek P, Bonfield JK, Liddle J et al (2021) Twelve years of samtools and bcftools. GigaScience 10

[125] Liao Y, Smyth GK, Shi W (2019) The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads. Nucleic Acids Research 47

[126] Kim D, Langmead B, Salzberg SL (2015) Hisat: a fast spliced aligner with low memory requirements. Nature Methods 12:357–360

[127] Pek JW, Okamura K (2015) Regulatory rnas discovered in unexpected places. Wiley Interdisciplinary Reviews: RNA 6:671–686

[128] Zhang Y, Xue W, Li X et al (2016) The biogenesis of nascent circular rnas. Cell Reports 15:611–624

[129] Zheng Y, Zhao F, Zhao F (2020) Visualization of circular rnas and their internal splicing events from transcriptomic data. Bioinformatics 36:2934–2935

[130] Paci P, Fiscon G, Conte F et al (2021) Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. npj Systems Biology and Applications 7

[131] Bain G, Ray WJ, Yao M, Gottlieb DI (1996) Retinoic acid promotes neural and represses mesodermal gene expression in mouse embryonic stem cells in culture. BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS 223:691–694

[132] Tagliaferri D, Mazzone P, Noviello TM et al (2020) Retinoic acid induces embryonic stem cells (escs) transition to 2 cell-like state

through a coordinated expression of dux and duxbl1. Frontiers in Cell and Developmental Biology 7

[133] Angelis MTD, Parrotta EI, Santamaria G, Cuda G (2018) Short-term retinoic acid treatment sustains pluripotency and suppresses differentiation of human induced pluripotent stem cells. Cell Death and Disease 9

[134] Miao S, Zhao D, Wang X et al (2020) Retinoic acid promotes metabolic maturation of human embryonic stem cell-derived cardiomyocytes. Theranostics 10:9686–9701

[135] Zhang J, Gao Y, Yu M et al (2015) Retinoic acid induces embryonic stem cell differentiation by altering both encoding rna and microrna expression. PLoS ONE 10

[136] Huang L, Chen M, Zhang W et al (2018) Retinoid acid and taurine promote neurod1-induced differentiation of induced pluripotent stem cells into retinal ganglion cells. Molecular and Cellular Biochemistry 438:67–76

[137] Mezquita B, Mezquita C (2019) Two opposing faces of retinoic acid: Induction of stemness or induction of differentiation depending on cell-type. Biomolecules 9

[138] Takahashi K, Nakamura M, Okubo C et al (2021) The pluripotent stem cell-specific transcript esrg is dispensable for human pluripotency. PLoS Genetics 17

[139] Wang J, Xie G, Singh M et al (2014) Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature 516:405–409

[140] Stefanov SR, Meyer IM *Deciphering the Universe of RNA Structures and trans RNA – RNA Interactions of Transcriptomes In Vivo : From Experimental Protocols to Computational Analyses* Springer (2018)

[141] Lipchina I, Elkabetz Y, Hafner M et al (2011) Genome-wide identification of microrna targets in human es cells reveals a role for mir-302 in modulating bmp response. Genes and Development 25:2173–2186

[142] S A Fastqc. (2015)

[143] Picard toolkit. https://broadinstitute.github.io/picard/ (2019)

[144] Zhang T, Zheng R, Li M et al (2022) Active endogenous retroviral elements in human pluripotent stem cells play a role in regulating host gene expression. Nucleic Acids Research 50:4959–4973

[145] Djebali S, Davis CA, Merkel A et al (2012) Landscape of transcription in human cells. Nature 489:101–108

[146] Cenik C, Cenik ES, Byeon GW et al (2015) Integrative analysis of rna, translation, and protein levels reveals distinct regulatory variation across humans. Genome Research 25:1610–1621

[147] A A, J R topgo: Enrichment analysis for gene ontology. r package version 2.50.0. https://bioconductor.org/packages/release/bioc/html/topGO.html (2022)

[148] Wenzel A, Akbaşli E, Gorodkin J (2012) Risearch: Fast rna-rna interaction search using a simplified nearest-neighbor energy model. Bioinformatics 28:2738–2746

[149] Li H, Durbin R (2010) Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics 26:589–595

[150] Hacisuleyman E, Goff LA, Trapnell C et al (2014) Topological organization of multichromosomal regions by the long intergenic noncoding rna firre. Nature Structural and Molecular Biology 21:198–206

[151] Chung H, Calis JJA, Wu X et al (2018) Human adar1 prevents endogenous rna from triggering translational shutdown in brief hhs public access the human rna editing enzyme adar1 prevents endogenous rna from activating innate immune sensors (pkr, mda5), which allows efficient translation during t. Cell 172:811–824

[152] Hofacker IL, Stadler PF (2006) Memory efficient folding algorithms for circular rna secondary structures. Bioinformatics 22:1172–1176

[153] Ahmad S, Mu X, Yang F et al (2018) Breaching self-tolerance to alu duplex rna underlies mda5-mediated inflammation. Cell 172:797–810.e13

[154] Gao Z, Zhu X, Dou Y (2015) The mir-302/367 cluster: A comprehensive update on its evolution and functions. Open Biology 5

Chapter 7

Appendices

7.1 Executing CYCLeR

Working with CYCLeR

Stefan Stefanov

9/14/2022

Contents

CYCLeR is a pipeline for reconstruction of circRNA transcripts from RNA-seq data and their subsequent quantification. The algorithm relies on comparison between control total RNA-seq samples and circRNA enriched samples to identify circRNA specific features. Then the selected circRNA features are used to infer the transcripts through a graph-based algorithm. Once the predicted transcript set is assembled, the transcript abundances are estimated through an EM algorithm with **kallisto** [1]. **CYCLeR** takes as an input BAM files and back-splice junction (BSJ) files and outputs transcript infomation in different formats and a transcript abundance file.

Installation of CYCLeR

Command line tools needed

The computation steps prior and post **CYCLeR** run are most efficiently run on HPC. It is very likely that any HPC in biological institute already has most of those tools installed. Just in case, a **Docker** image containing all the tools is provided.
NOTE: prior to running **Docker** image make sure that ***Docker** is indeed installed and working: https://docs.docker.com/get-started/

- **STAR** - https://github.com/alexdobin/STAR
- **samtools** - https://sourceforge.net/projects/samtools/files/samtools/
- **kallisto** - http://pachterlab.github.io/kallisto/download
- **bwa** (needed for CIRI2) - http://bio-bwa.sourceforge.net/bwa.shtml
- **CIRI2** - https://sourceforge.net/projects/ciri/files/CIRI2/
- **CIRCexplorer2** - https://circexplorer2.readthedocs.io/en/latest/

```
#Docker image with all command line tools
sudo docker pull stiv1n/cycler.prerequisites
```

R packages installation

Option 1: Local installation from GitHub Standard GitHub installation. The dependencies might have compilation issues. For Ubuntu, the issues should be resolved with installation of a few libraries. NOTE: you need a local **samtools** binary for an optional step.

```
#sudo apt update && apt install -y libcurl4-openssl-dev libxml2 libssl-dev \
#libbz2-dev liblzma-dev pkg-config build-essential libglpk40
library(devtools)
install_github("stiv1n/CYCLeR")
```

Option 2: Docker installation The **Docker** use requires you to mount a volume - a working directory () where the output and input would be stored. This container uses **RStudio server** and required login. In this case, the username is *rstudio* the password is *guest*.

```
sudo docker pull stiv1n/cycler
sudo docker run --rm -ti -e PASSWORD=guest -v <local_dir>:/usr/workdir -p 8787:8787 stiv1n/cycler
```

Pre-processing the data

Mapping with STAR

The STAR [2] mapping parameters are up to a personal preference. It is imperative to include the **intronMotif** tag. Sorting of the file can be performed via **STAR** or **samtools**. NOTE: STAR requires an index and works better with provided GTF. The parameters of STAR index are dependent on the sequencing, so better to read the manual.
An example run for the **Docker** container is shown for **samtools**. My preferred parameters for **STAR**:

```
#STAR parameters
STAR --alignSJoverhangMin 8 --outSAMstrandField intronMotif
--outFilterMismatchNmax 2 --outFilterMismatchNoverLmax 0.1 --chimSegmentMin 15
--chimScoreMin 1 --chimJunctionOverhangMin 15 --chimOutType WithinBAM
--outSAMtype BAM SortedByCoordinate --limitBAMsortRAM 9664623958
--outFilterMultimapNmax 50 --alignIntronMax 100000 --alignIntronMin 15
--seedSearchStartLmax 5 --winAnchorMultimapNmax 200
#Docker run
sudo docker run -v <local_dir>:/usr/local stiv1n/cycler.prerequisites STAR
```

Again, the **Docker** use requires you to mount a volume - a working directory () where the output and input would be stored.

```
#converting the default Aligned.out.sam to a sorted BAM
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  samtools view -u -h /usr/local/Aligned.out.sam | samtools sort \
  -o /usr/local/<name>_sorted.bam
```

BSJ identification

It is advantageous to have input from BSJ identification tools that use different aligners. I suggest **CIRI2** with **bwa** and **CIRCexplorer2** with **STAR**. We have already discussed **STAR** mapping. **CIRI2** requires **bwa** mapping. NOTE: For safety always use full path.

```
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  bwa index -a bwtsw reference.fa
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  bwa mem -T 19 reference.fa read_1.fq read_2.fq > <sample_name>.sam
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  perl /usr/src/myapp/CIRI_v2.0.6/CIRI2.pl -I <sample_name>.sam \
  -O CIRI_<sample_name> -F reference.fa -A annotation.gtf
```

CYCLeR needs just 2 steps of the **CIRCexplorer2** pipeline. NOTE: **CIRCexplore2** uses a flat annotation file

```
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  CIRCexplorer2 parse -t STAR /usr/local/Chimeric.out.junction \
  -b /usr/local/back_spliced_junction.bed
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  CIRCexplorer2 annotate -r annotation.txt -g reference.fa \
  -b /usr/local/back_spliced_junction.bed -o /usr/local/CE_<sample_name>
```

CYCLeR

After all the pre-processing, all the files should preferably be in one folder.

Processing the BAM info in R

We need the information for read length, fragment length and library sizes from the BAM files.

```
#load BAM files
bam_file_prefix<-system.file("extdata", package = "CYCLeR")
filenames<-c("sample1_75","sample2_75","sample3_75","sample4_75")
BSJ_files_ciri<-paste0(bam_file_prefix,"/",filenames)
bam_files<-paste0(bam_file_prefix,"/",filenames,".bam")
#mark the samples control and enriched or bare the consequences
sample_table<-data.frame(filenames,c("control","control","enriched","enriched")
                         ,bam_files,stringsAsFactors = F)
colnames(sample_table)<-c("sample_name","treatment","file_bam")
si<- DataFrame(sample_table[,c("sample_name","file_bam")])
si$file_bam <-BamFileList(si$file_bam, asMates = F)
#this holds all the needed info of the bam files for downstream processing
sc <- getBamInfo(si)
sample_table$lib_size<-sc@listData$lib_size
sample_table$read_len<-sc@listData$read_length
```

Use the provided sample table template.

```
##   sample_name treatment
## 1  sample1_75   control
## 2  sample2_75   control
## 3  sample3_75  enriched
## 4  sample4_75  enriched
##                                                                           file_bam
## 1 /home/stefan/miniconda3/envs/cycler/lib/R/library/CYCLeR/extdata/sample1_75.bam
## 2 /home/stefan/miniconda3/envs/cycler/lib/R/library/CYCLeR/extdata/sample2_75.bam
## 3 /home/stefan/miniconda3/envs/cycler/lib/R/library/CYCLeR/extdata/sample3_75.bam
## 4 /home/stefan/miniconda3/envs/cycler/lib/R/library/CYCLeR/extdata/sample4_75.bam
##   lib_size read_len
## 1    13884       75
## 2    13959        75
## 3     8494        75
## 4     8637        75
```

Selecting a BSJ set

Selecting a BSJ set is very important, because the algorithm assumes that the provided set of BSJ is *correct*. I suggest BSJ identification with **CIRI2** [3] and **CIRCexplorer2** [4], but the choice is up to a personal preference. I have provided some useful functions for parsing the output from BSJ identification software.

```
#load the BSJ files
BSJ_files_prefix<-paste0(system.file("extdata", package = "CYCLeR"),"/ciri_")
ciri_table<-parse_files(sample_table$sample_name,BSJ_files_prefix,"CIRI")
colnames(ciri_table)<-c("circ_id", "sample1_75","sample2_75","sample3_75","sample4_75")
ciri_bsjs<-process_BSJs(ciri_table,sample_table)
# i would suggest combine the output of pipelines using different mapping tools
BSJ_files_prefix_CE<-paste0(system.file("extdata", package = "CYCLeR"),"/CE_")
ce_table<-parse_files(sample_table$sample_name,BSJ_files_prefix_CE,"CE")
colnames(ce_table)<-c("circ_id", "sample1_75","sample2_75","sample3_75","sample4_75")
ce_bsjs<-process_BSJs(ce_table,sample_table)
#we need to unify the results from the BSJ identification and counting
table_circ<-combine_two_BSJ_tables(ce_bsjs,ciri_bsjs)
#further downstream we need just the mean values for enriched samples
table_circ<-table_circ[,c("chr","start","end","meanRR")]
colnames(table_circ)<-c("chr","start","end","count")
#combine
BSJ_set<-union(ciri_bsjs$circ_id,ce_bsjs$circ_id)
BSJ_set<-BSJ_set[!grepl("caffold",BSJ_set)]
#just in case
BSJ_set<-BSJ_set[!grepl("mitochondrion",BSJ_set)]
################################################################
#converting the BSJ set into a GRanges object
BSJ_gr<-make_BSJ_gr(BSJ_set)
```

The parse.files can work with **CIRI2**, **CIRCexplorer2** or **TSV** file. Naturally a person may have different criterion for *correct* BSJs based on different criteria. It is not an issue as long as the data is presented in the following template:

```
head(table_circ)
```

```
## # A tibble: 6 x 4
##   chr   start    end        count
##   <chr> <chr>    <chr>      <dbl>
## 1 3L    24725824 24726292 435115.
## 2 3L    24725824 24728508   9150.
## 3 3L    24728297 24734187 171930.
## 4 3L    24728297 24741000   7992.
## 5 3R    4622509  4628349   64579.
## 6 3R    4626973  4628349  160551.
```

If you use `parse_files` with `input_type=="tsv"`, you can just edit the table with:

```
table_circ<-table_circ%>%separate(circ_id, into=c("chr","start","end","strand"),sep = "_")
```

Transcript assembly

BSJ loci extraction (optional) Prior to the feature detection the files need to be trimmed to speed up the process. Afterwards the transcript features (e.g. exons, junctions) are identified with **SGSeq** [5]. The files are processed with **samtools** [6]. The trimmed files are also useful for long term local storage.

```
####################################################
samtools_prefix<-""
trimmed_bams<-filter_bam(BSJ_gr,sample_table,samtools_prefix)
sc@listData[["file_bam"]]<-trimmed_bams
```

Annotation info (optional) The use of the TxDb package is to annotate the identified features. The annotation step is not mandatory, but it does provide useful information. I can also be used to avoid *de novo* feature detection. In the **Docker** container the annotation library is provided.

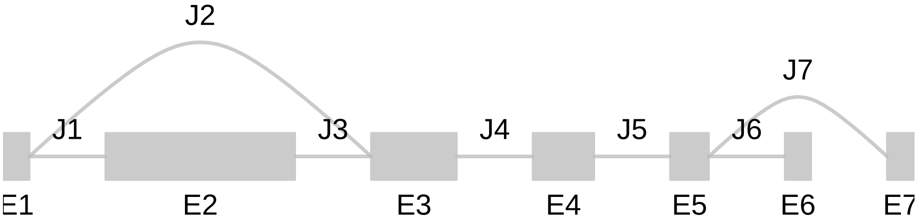
```
#####################################################
#get the gene/transcript info
library("TxDb.Dmelanogaster.UCSC.dm6.ensGene")
#restoreSeqlevels(txdb)
txdb <- TxDb.Dmelanogaster.UCSC.dm6.ensGene
txdb <- keepSeqlevels(txdb, c("chr2L","chr2R","chr3R","chr3L","chr4","chrX","chrY"))
seqlevelsStyle(txdb) <- "Ensembl"
gene_ranges <- genes(txdb)
txf <- convertToTxFeatures(txdb)
#asnnotation as sg-object
sgf <- convertToSGFeatures(txf)
```

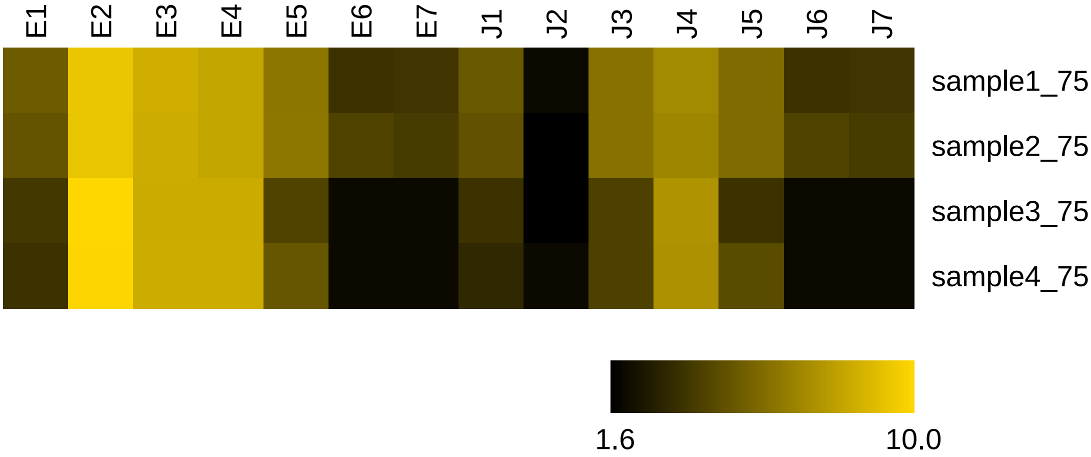
Feature identification with SGSeq The feature detection is based on the SGSeq package. There are 3 options to approach the problem. The default function parameters requires a lot of RAM and processing time. The second option allows using **Rsamtools** `which=BSJ_gr` function, which focuses the reconstruction solely on the pre-selected regions. This significantly speeds up the feature detection and lowers RAM requirements. It is less reliable than the default parameters, however, it is very convenient for a quick test. Additionally, `features=txf` can be provided to indicate that no *de novo* assembly should be done. Naturally, that is the fastest approach, but obviously lacking.

```
#option 1: for fast computer, no RAM limitations, time is not a factor
sgfc_pred <- analyzeFeatures(sc, min_junction_count=2, beta =0.1 ,
                             min_n_sample=1,cores=1,verbose=F)

#option 2: for moderate computer, limited RAM, speed is of the essence
sgfc_pred <- analyzeFeatures(sc, which=BSJ_gr, min_junction_count=2, beta =0.1 ,
                             min_n_sample=1,cores=1,verbose=F)

#option 3: for a toaster with attached monitor
sgfc_pred <- analyzeFeatures(sc, which=BSJ_gr, features=txf, min_junction_count=2,
                             beta =0.1 , min_n_sample=1,cores=1,verbose=F)
#annotation (optional)
sgfc_pred <- SGSeq::annotate(sgfc_pred, txf)
```

SGSeq feature plotting function can be used for visual representation of the control VS enriched difference

```
plotFeatures(sgfc_pred,  geneID = "1",assay = "counts",color_novel = "red",
             include = "both",tx_view=F,Rowv=NA, square=T)
```


J2
J7
J1
J3
J4
J5
J6
E1
E2
E3
E4
E5
E6
E7


E1
E2
E3
E4
E5
E6
E7
J1
J2
J3
J4
J5
J6
J7
sample1_75
sample2_75
sample3_75
sample4_75
1.6
10.0

Re-couting and Processing the features

I prefer the **RSubread** [7] counting method, thus I re-count the identified exon features. Later the **SGseq** counted junctions are used. The features that are depleted in circRNA enriched samples need to be removed. **CYCLeR** provides 2 approaches for identifying depleted features: DEU strategy and simple comparison of normalized coverage values. The simple comparison turns on automatically only due to the lack of replicates.

```
#extract BSJ-corrected splice graphs (sg)
full_sg<-overlap_SG_BSJ(sgfc_pred,BSJ_gr,sgf) #includes linear and circular features
# we have made new feature set so we need to recount the exons
full_fc<-recount_features(full_sg,sample_table)#fc==feature counts
# time to prepare the circular splice graph
```

The sequences of the exons are needed for the subsequent steps. The genome sequence is provided with a **BSgenome** package. For the tutorial the **Docker** image has the needed library provided.

```
#get the correct genome for sequence info
#requires the appropriate BSgenome library
library(BSgenome.Dmelanogaster.UCSC.dm6)
bs_genome=Dmelanogaster
circ_sgfc<-prep_circular_sg(full_sg, full_fc,sgfc_pred, bs_genome, BSJ_gr, th=15)
```

The circRNA exon features are stored in SummarizedExperiment format

Transcript prediction

Transcript prediction is processed one samples at a time. The transcript sets from different samples are then merged.

```
qics_out1<-transcripts_per_sample(sgfc=circ_sgfc,BSJ_gr = BSJ_gr,"sample3_75")
qics_out2<-transcripts_per_sample(sgfc=circ_sgfc,BSJ_gr = BSJ_gr,"sample4_75")
qics_out_final<-merge_qics(qics_out1,qics_out2,sgfc_pred)
```

Output and Quantification

CYCLeR transcript output

CYCLeR provides 3 forms of output of the annotated transcript: a comprehensive flat file, a GTF-like file, and FASTA file.

```
gtf.table<-prep_output_gtf(qics_out_final,circ_sgfc)
write.table(qics_out_final[,-9],file = "dm_circles.txt", sep = "\t",
            row.names = F, col.names = T,quote=F)
qics_out_fa<-DNAStringSet(qics_out_final$seq)
names(qics_out_fa)<-qics_out_final$circID

#prepping the circRNA sequences for quantification
extended_seq<-paste0(qics_out_final$seq,substr(qics_out_final$seq,1,30),
      strrep("N",mean(sc@listData$frag_length[sample_table$treatment=="enriched"])))
qics_out_fa_extended<-DNAStringSet(extended_seq)
names(qics_out_fa_extended)<-qics_out_final$circID
writeXStringSet(qics_out_fa_extended,'circles_seq_extended_padded.fa')
```

If you have a known set of circRNA in FASTA format the CYCLeR output can be combined with it.

```
fasta_circ<-readDNAStringSet("...")
final_ref_fa<-merge_fasta(qics_out_fa,fasta_circ)
writeXStringSet(final_ref_fa,'...')
```

The same function can be used for merging with known linear transcript sequences for the quantification step.

```
fasta_lin<-readDNAStringSet("...")
final_ref_fa<-merge_fasta(qics_out_fa_extended,fasta_lin)
writeXStringSet(final_ref_fa,'for_kallisto.fa')
```

CYCLeR quantification

The final transcript abundance estimation is performed with **kallisto**. An extended and padded circRNA reference sequences are build and combined with linear RNA sequences *kallisto index* is created to be used for any desired sample quantification.

```
#alternative way of merging linear and circular sequences
cat linear_transcripts.fa circles_seq_extended_padded.fa > for_kallisto.fa
#Kallisto comands
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  kallisto index -i kallisto_index -k 31 for_kallisto.fa
sudo docker run  -v <local_dir>:/usr/local stiv1n/cycler.prerequisites \
  kallisto quant -i kallisto_index -o ./ <sample_name>_1.fastq <sample_name>_2.fastq
```

1. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;34:525–7.
2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
3. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. Briefings in bioinformatics. 2018;19:803–10.
4. Zhang X, Dong R, Zhang Y, Zhang J, Luo Z, Zhang J, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Research. 2016;1277–87.
5. Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, et al. Prediction and quantification of splice events from RNA-seq data. PLoS ONE. 2016;11:1–8.
6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
7. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research. 2019;47.

7.2 Supplementary tables and figures

Organism	Reference genome	Annotation	Source
Fruit fly	BDGP6 (dm6)	BDGP6.87	Ensembl Top level Assembly
Mouse	GRCm39 (mm39)	GRCm39.104	Ensembl Primary Assembly
Human	GRCh38 (hs38)	GRCh38.101	Ensembl Top level Assembly

Table 7.1: **Summary of reference genome and annotation versions used in the work for Chapter3.** *The source of all files is the Ensembl FTP server.* Note: *for* CYCLeR *feature annotation, the corresponding* R TxDb *package was used.*

Organism	Reference genome	Annotation	Source
Mouse	GRCm38 (mm10)	GRCm38.101	Ensembl Top level Assembly
Human	GRCh38 (hs38)	GRCh38.101	Ensembl Top level Assembly

Table 7.2: **Summary of reference genome and annotation versions used in the work for Chapter4.** *The source of all files is the Ensembl FTP server.* Note: *for* CYCLeR *feature annotation, the corresponding* R TxDb *package was used.*

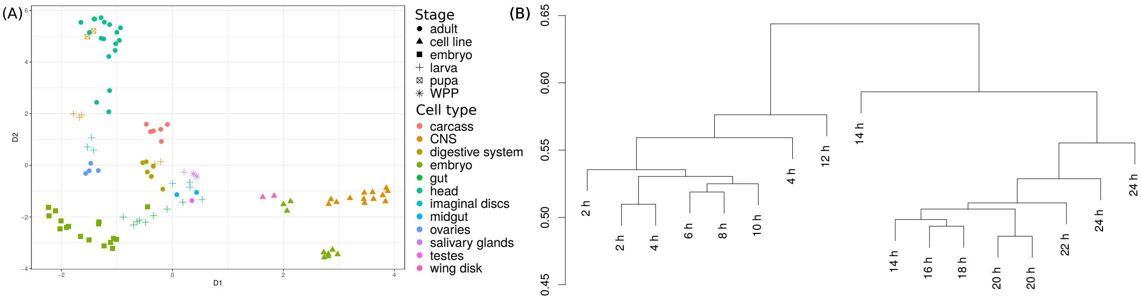


Figure 7.1: **Sailfish-cir applied to Lai lab 2014 dataset.** *Extension from the corresponding figure in the manuscript. In part (A), is shown the UMAP-dimensional scaling of the abundances generated by sailfish-cir of all 103 samples from the dataset. In part (B), we use the abundances computed by sailfish-cir to plot a dendrogram of the embryo stages subset based on between-sample distance calculations. sailfish-cir has the same quantification strategy as* CYCLeR, *but lacks a proper assembly algorithm. The results indicate that the quantification strategy itself is not sufficient for improved clustering and the correct sequences of the isoforms are imperative for correct quantification.* ***Reproduced from Stefanov et al. (2022)***

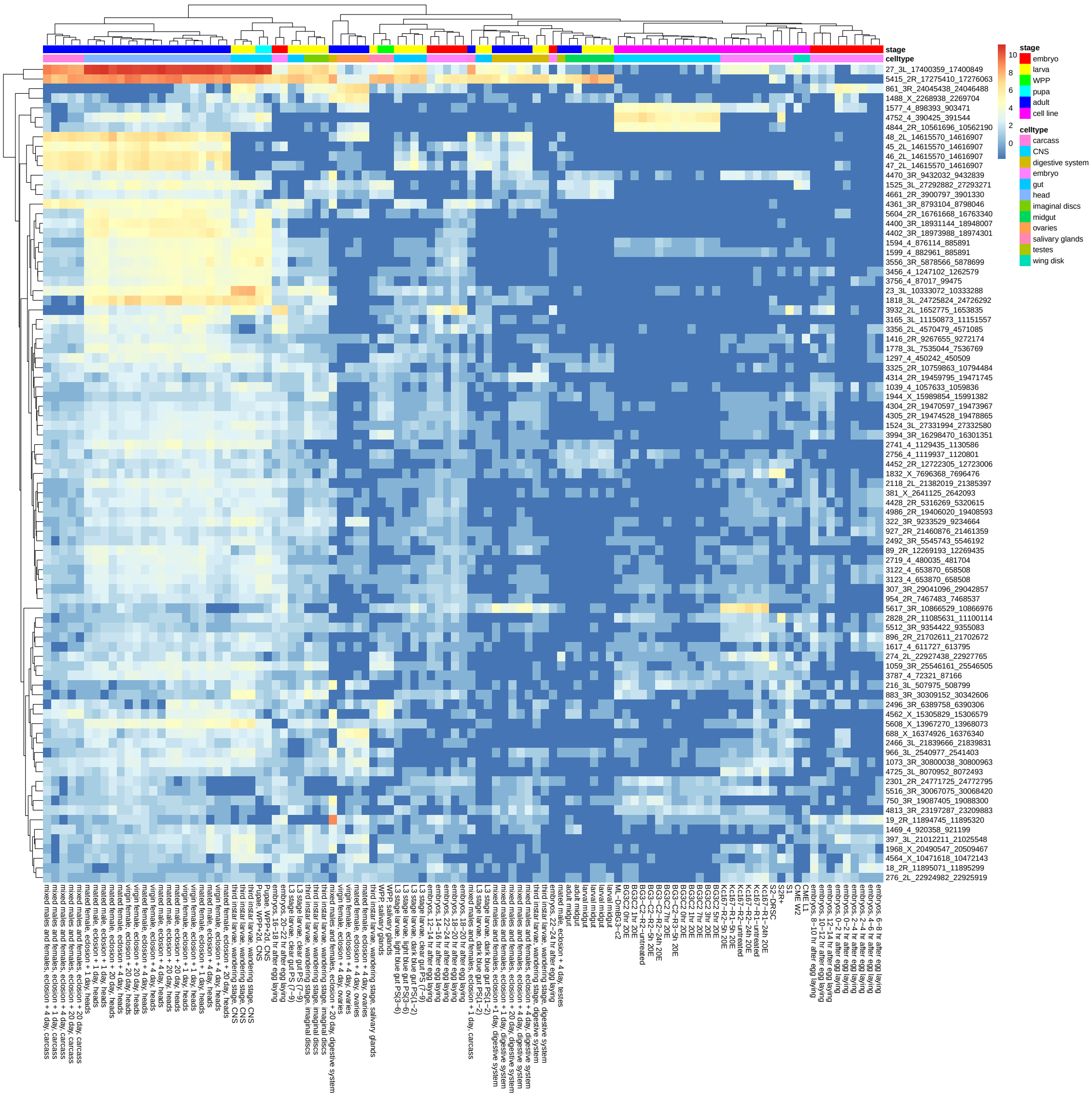


Figure 7.2: **Heatmap of most variable circRNAs** *for the D. Lai dataset.* CYCLeR *can identify circRNAs that are specific for a particular stage of development or a particular cell type. Furthermore, it can differentiate circRNAs specific for the CNS or circRNAs specific for adult flies. We can also conclude that embryo derived cell lines show a similar pattern as early stage embryo samples.* ***Reproduced from Stefanov et al. (2022)***

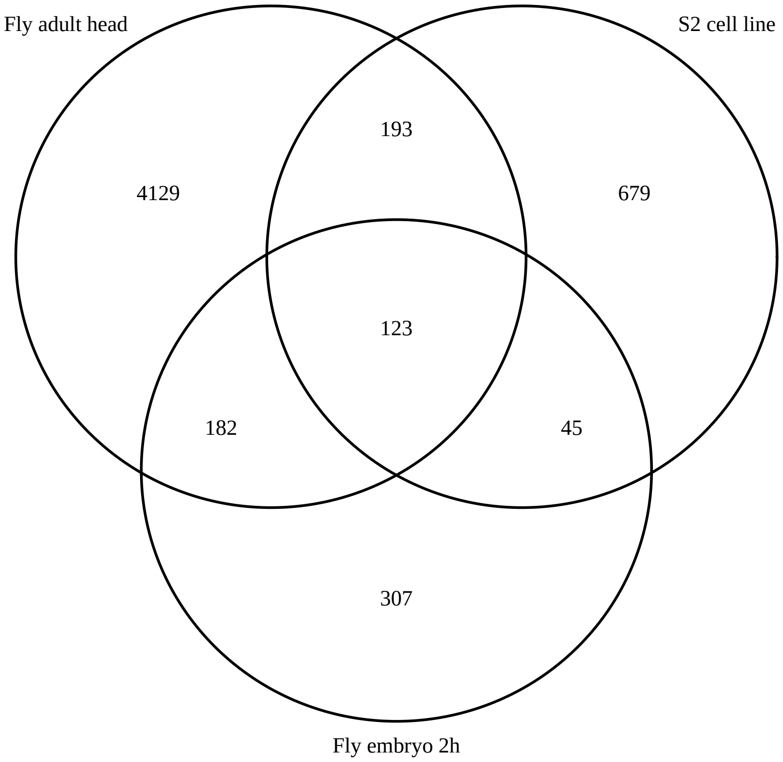


Figure 7.3: **Fly transcripts overlap** *Overlap between the sets of transcripts that are assembled by* CYCLeR*. The data from replicates is merged into one set.**Reproduced from Stefanov et al. (2022)***

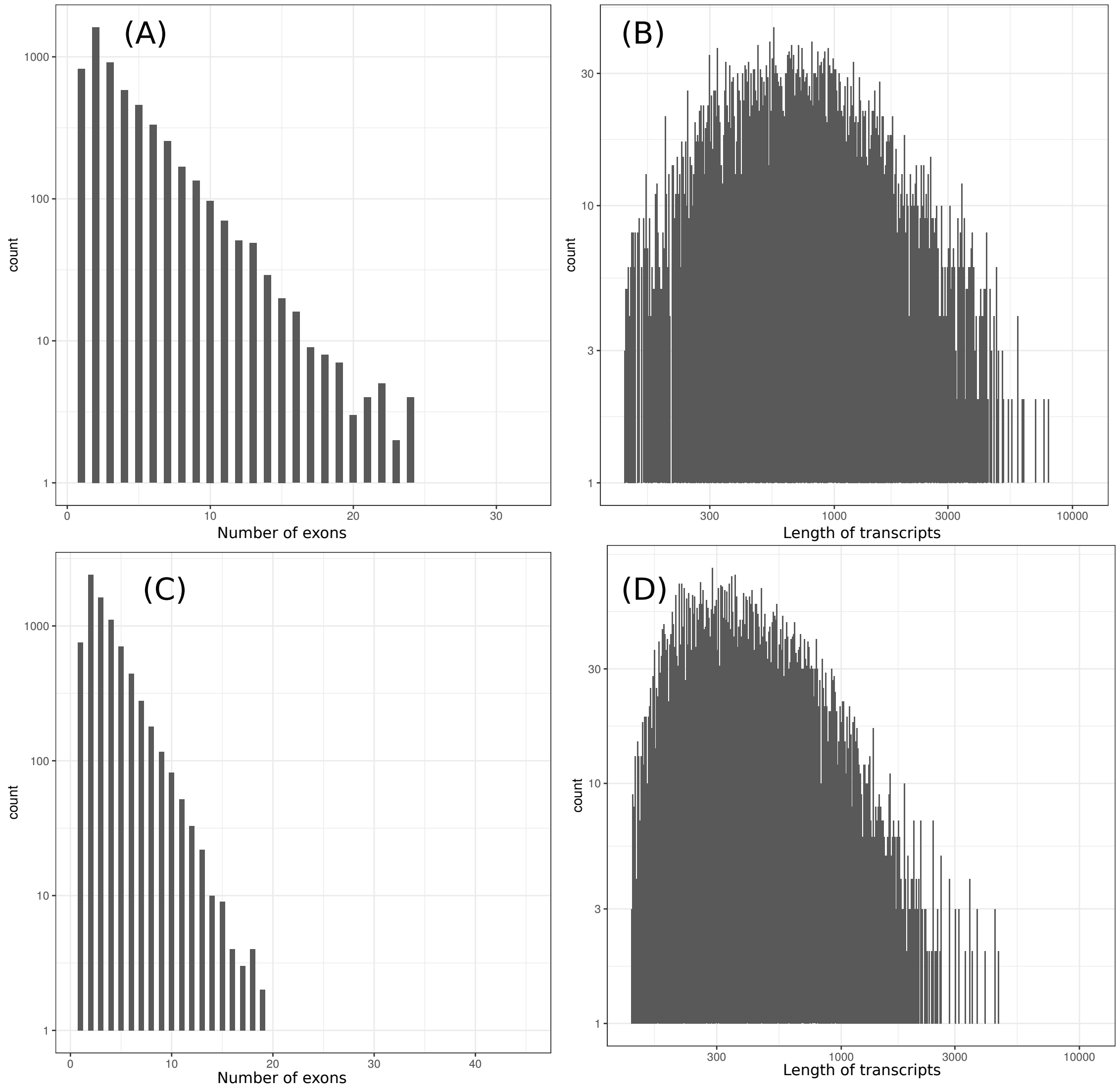
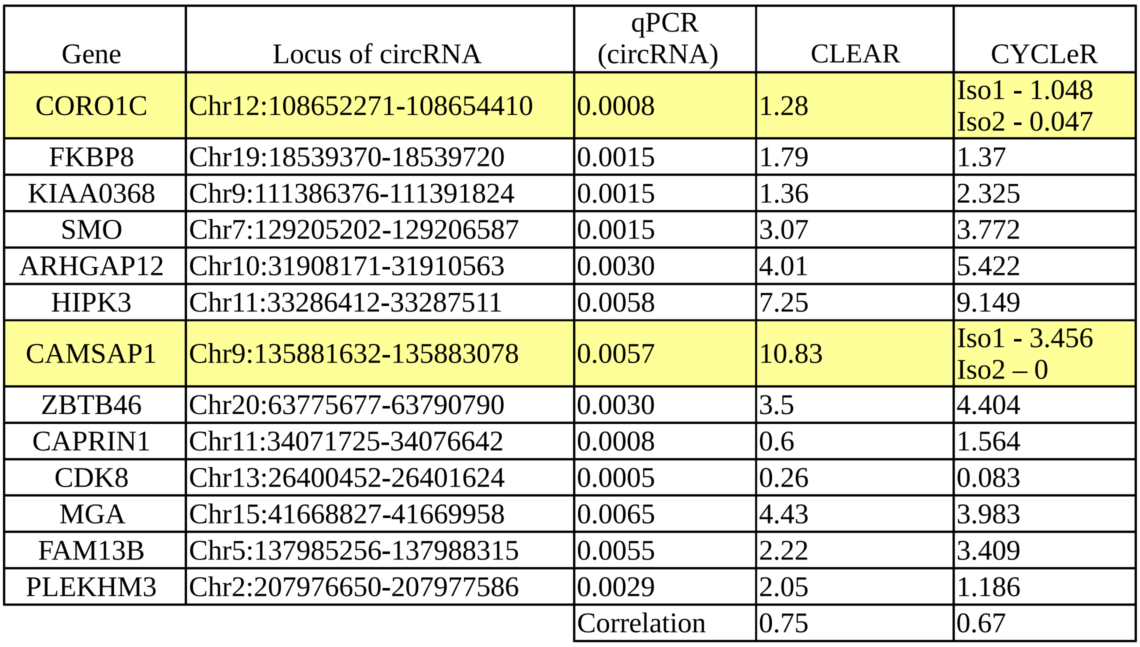
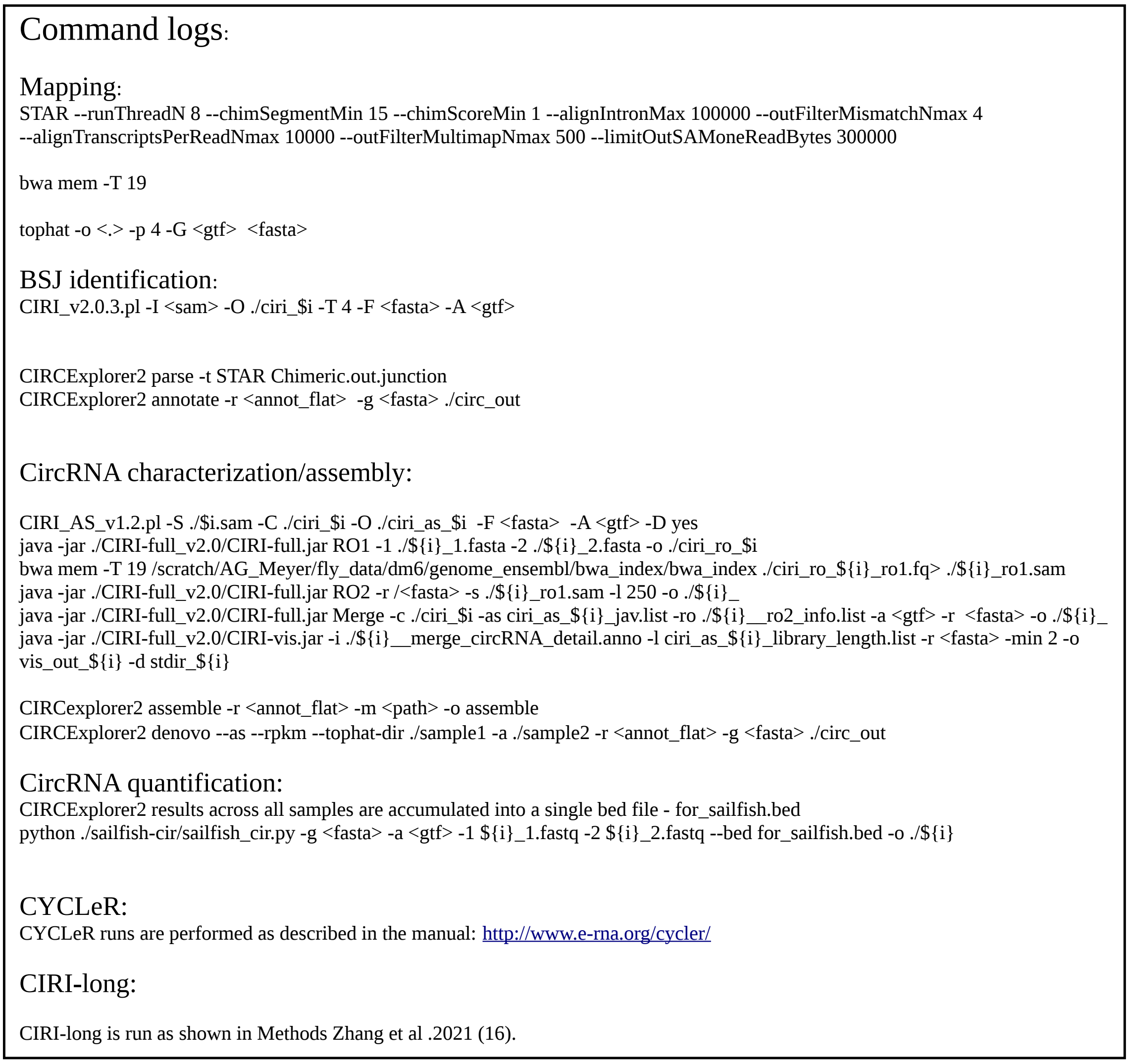


Figure 7.4: **Statistics of real data assembly.** *The results show the reconstruction of the PA1 dataset (A) and (B) and the accumulated data of the fruit fly dataset (C) and (D). (A) and (C) show the number of exons per transcript and (B) and (D) show the length of transcript on a logarithmic scale.* ***Reproduced from Stefanov et al. (2022)***



Gene	Locus of circRNA	qPCR (circRNA)	CLEAR	CYCLeR
CORO1C	Chr12:108652271-108654410	0.0008	1.28	Iso1 - 1.048 Iso2 - 0.047
FKBP8	Chr19:18539370-18539720	0.0015	1.79	1.37
KIAA0368	Chr9:111386376-111391824	0.0015	1.36	2.325
SMO	Chr7:129205202-129206587	0.0015	3.07	3.772
ARHGAP12	Chr10:31908171-31910563	0.0030	4.01	5.422
HIPK3	Chr11:33286412-33287511	0.0058	7.25	9.149
CAMSAP1	Chr9:135881632-135883078	0.0057	10.83	Iso1 - 3.456 Iso2 – 0
ZBTB46	Chr20:63775677-63790790	0.0030	3.5	4.404
CAPRIN1	Chr11:34071725-34076642	0.0008	0.6	1.564
CDK8	Chr13:26400452-26401624	0.0005	0.26	0.083
MGA	Chr15:41668827-41669958	0.0065	4.43	3.983
FAM13B	Chr5:137985256-137988315	0.0055	2.22	3.409
PLEKHM3	Chr2:207976650-207977586	0.0029	2.05	1.186
		Correlation	0.75	0.67

Figure 7.5: **Results for PA1 RNA-seq data.** *All values in the table correspond to averages between replicates. The yellow rows indicate the filtered items from the benchmark due to known multiple isoforms. The correlation cells show the Pearson product correlation of the filtered values of qPCR results and estimated abundances. Note that the CAMSAP1 Iso2 has zero value due to the fact that with default parameters,* CYCLeR *fails to recover the more abundant isoform. This is due to multiple overlapping circRNA isoforms, whose reconstruction leads to premature depletion of the reconstruction algorithm.* ***Reproduced from Stefanov et al. (2022)***



Command logs:

Mapping:

```
STAR --runThreadN 8 --chimSegmentMin 15 --chimScoreMin 1 --alignIntronMax 100000 --outFilterMismatchNmax 4
--alignTranscriptsPerReadNmax 10000 --outFilterMultimapNmax 500 --limitOutSAMoneReadBytes 300000

bwa mem -T 19

tophat -o <.> -p 4 -G <gtf>  <fasta>
```

BSJ identification:

```
CIRI_v2.0.3.pl -I <sam> -O ./ciri_$i -T 4 -F <fasta> -A <gtf>


CIRCExplorer2 parse -t STAR Chimeric.out.junction
CIRCExplorer2 annotate -r <annot_flat>  -g <fasta> ./circ_out
```

CircRNA characterization/assembly:

```
CIRI_AS_v1.2.pl -S ./$i.sam -C ./ciri_$i -O ./ciri_as_$i  -F <fasta>  -A <gtf> -D yes
java -jar ./CIRI-full_v2.0/CIRI-full.jar RO1 -1 ./${i}_1.fasta -2 ./${i}_2.fasta -o ./ciri_ro_$i
bwa mem -T 19 /scratch/AG_Meyer/fly_data/dm6/genome_ensembl/bwa_index/bwa_index ./ciri_ro_${i}_ro1.fq> ./${i}_ro1.sam
java -jar ./CIRI-full_v2.0/CIRI-full.jar RO2 -r /<fasta> -s ./${i}_ro1.sam -l 250 -o ./${i}_
java -jar ./CIRI-full_v2.0/CIRI-full.jar Merge -c ./ciri_$i -as ciri_as_${i}_jav.list -ro ./${i}__ro2_info.list -a <gtf> -r  <fasta> -o ./${i}_
java -jar ./CIRI-full_v2.0/CIRI-vis.jar -i ./${i}__merge_circRNA_detail.anno -l ciri_as_${i}_library_length.list -r <fasta> -min 2 -o
vis_out_${i} -d stdir_${i}

CIRCexplorer2 assemble -r <annot_flat> -m <path> -o assemble
CIRCExplorer2 denovo --as --rpkm --tophat-dir ./sample1 -a ./sample2 -r <annot_flat> -g <fasta> ./circ_out
```

CircRNA quantification:

CIRCExplorer2 results across all samples are accumulated into a single bed file - for_sailfish.bed

```
python ./sailfish-cir/sailfish_cir.py -g <fasta> -a <gtf> -1 ${i}_1.fastq -2 ${i}_2.fastq --bed for_sailfish.bed -o ./${i}
```

CYCLeR:

CYCLeR runs are performed as described in the manual: http://www.e-rna.org/cycler/

CIRI-long:

CIRI-long is run as shown in Methods Zhang et al .2021 (16).

Figure 7.6: **Overview of commands used** *Information about the annotation and parameters used for the tools [49, 50]. In this log, "i" stands for file prefix.*