Exon-intron chain reconstruction of circular RNA using RNA-Seq

Inaugural-Disseration
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Franziska Metge
aus Leipzig

Köln 2018

Gutachter: Dr. Dario R. Valenzano
Prof. Dr. Andreas Beyer

Tag der mündlichen Prüfung: 6. Oktober 2017

You never know ...

Olivia Lynn DeYonge

Kurzzusammenfassung

Motivation: Zirkuläre Ribonukleinsäuren (circRNA) sind eine spezifische Gruppe von RNA, welche einen kovalent geschlossenen Kreis bilden. Dieser Prozess wird als rückwärts-spleißen bezeichnet. Bisher ist nicht viel über die funktionsweise von circRNAs bekannt. Für einige wenige circRNAs wurden mögliche Funktionen als Schwamm für microRNAs (miRNA) oder RNA-bindende Proteine (RBP) gezeigt. Außerdem können sie die Transkription ihres Wirts-Gens regulieren. Zirkuläre RNA können anhand von chimären Sequenzen, welche die rückwärts gespleißte Verbindung überbrücken, in rRNA-verdauten RNA-Seq Bibliotheken detektiert werden. Zur Zeit gibt es eine Vielfalt an unterschiedlichen Programmen, die circRNAs in RNA-Seq Daten identifizieren. Allerdings ist keines der Programme in der Lage die circRNAs weiter zu characterisieren oder zusammenzufassen. Um weiterführende Analysen an den entdeckten circRNAs durch führen zu können ist es unablässlich die genaue Exon-Intron Struktur von circRNAs zu kennen. Vor kurzem wurden zwei neue Programme veröffentlicht, die alternatives Spleißen in circRNAs beschreiben. In meiner Arbeit stelle ich `FUCHS` und `FUCHS`*denovo* vor um entdeckte circRNAs zusammenzufassen und die Exon-Intron Struktur anhand von linearen Spleiß Signalen von chimären Sequenzen zu rekonstruieren.

Methoden: Zuerst habe ich drei der aktuellsten circRNA Identifikations Progamme miteinander verglichen, um basierend auf den Ergebnissen des besten Programms eine Pipeline in Python zu entwicklen. Diese Pipeline heißt `FUCHS`, kurz für "**FU**ll **CH**aracterization of circular RNA using RNA-**S**equencing". Sie fasst circRNAs nach ihren Wirts-Genen zusammen, findet übersprungene Exons, findet doppeltchimäre Sequenzen, generiert Abdeckungsprofile und fasst circRNAs basierend auf ihrem Abdeckungsprofil zusammen. Das Anwenden von `FUCHS` auf einem Beispieldatensatz hat gezeigt, dass annotierte Strukturen oft nicht ausreichen um die zirkulären Strukturen zu beschreiben. Deswegen habe ich `FUCHS` erweitert. Das neue Programm heißt `FUCHS`*denovo*, da es die Exon-Intron Strukturen von circRNAs *de novo* rekonstruieren kann. Um die Funktionsweise beider Programme vorzustellen, habe ich sie auf einem Beispieldatensatz bestehend aus Leber und Herz Proben von jungen und alten Mäusen angewendet.

Ergebnisse: Im Vergleich von drei circRNA Identifikations Progammen (`DCC`, `CIRI`, and `KNIFE`), hob sich `DCC` als schnellstes und präzisestes Program ab. Die Anwendung von `FUCHS` auf vier Maus Proben zeigte, dass es weniger unterschiedliche circRNAs im Herzen als in der Leber gibt, diese dafür aber in höherer Anzahl. Betrachtet man nur annotierte Exons, zeigt sich, dass die circRNAs im Herzen länger sind als die in der Leber. Die durchschnittliche Länge der circRNAs beträgt 500 BP. Aus den Abdeckungsprofilen habe ich geschlossen, dass die annotierten Exon-Intron Strukturen nicht

immer mit den Exon-Intron Strukturen der circRNAs übereinstimmen. Ein Vergleich zwischen den annotierten und den mit FUCHS*denovo* rekonstruierten Strukturen zeigte einen Gewinn von 15 % an zusätzlicher Information. Weiterhin hat FUCHS*denovo* alternatives Spleißen in 8 - 10 % der circRNAs finden können. Eine Analyse von differenziel angereicherten Motifen in den Introns um circRNAs zeigte, dass die Introns um circRNAs mit alternativen Spleiß Isoformen gegenüber circRNAs ohne alternativen Spleiß Isoformen mit FOXO Bindemotifen angereichert sind. Bindemotife für CPEB1 und HOX waren in Introns um circRNAs von multi-circRNA Genen gegenüber Introns um circRNAs von single-circRNA Genen angereichert. Somit könnten sowohl FOXO als auch CPEB1 und HOX eine Rolle in der Biogenese von circRNAs spielen. Eine miRNA und RBP Bindemotif Suche hat gezeigt, dass Exons von circRNAs dichter mit Bindemotifen bestückt sind als Exons von linearen mRNAs. Daraus schließe ich, dass circRNAs eine weitere Ebene im Genregulationsnetzwerk darstellen können, indem sie mit linearen mRNAs für die Bindung von miRNAs und RBPs konkurrieren.
Verfügbarkeit: https://github.com/dieterich-lab/FUCHS.git

Abstract

Motivation: Circular RNAs (circRNAs) are a special class of RNA forming a covalently closed loop through a process called back-splicing. Not much is known about the function of circRNAs. Only for a few well studied circRNAs, potential functions were shown, these include miRNA sponging, RNA binding protein (RBP) sponging, and regulation of their host gene's transcription. Circular RNAs can be identified in rRNA depleted RNA-Sequencing by detecting chimeric reads, which span a back-splice junction. A variety of circRNA detection tools exists but no tool is able to summarize and characterize the identified circRNAs. To perform accurate downstream analyses after circRNA detection, it is crucial to know the exact exon-intron structure of circRNAs. Recently, two tools were published, which identify alternative splicing within circRNAs. Here, I am presenting `FUCHS` and `FUCHS`*denovo* to summarize circRNAs and reconstruct their exon-intron chain based on linear-splice signals of back-splice junction anchored reads.

Methods: In this study, I compared three state of the art circRNA detection programs. Based on the best tool, I developed a Python-based pipeline called `FUCHS`: **FU**ll **CH**aracterization of circular RNA using RNA-**S**equencing. This pipeline summarizes circRNAs by their host genes, detects skipped exons, finds double-breakpoint fragments, generates circle-wise coverage profiles, and clusters these profiles. Running `FUCHS` on a mouse dataset indicated that annotated gene models are not always suited to describe the circRNA's exon-intron structure. Hence, I developed an additional module, `FUCHS`*denovo*, to reconstruct the exon-intron structure based on linear-splice signals of back-splice junction anchored reads. To demonstrate how `FUCHS` and `FUCHS`*denovo* perform, I ran both programs on a dataset of young and old murine hearts and young and old murine livers.

Results: The comparison of three circRNA detection programs (`DCC`, `CIRI`, and `KNIFE`) indicated `DCC` as the fastest and most accurate circRNA detection program. Running `FUCHS` on four mouse samples revealed that heart circRNAs are less diverse but more abundant than liver circRNAs. Considering only annotated exons, the average length of circRNAs was 500 BP. Heart circRNAs were longer than liver circRNAs. From the obtained coverage profiles, I concluded that annotated gene models were not always matching the exon-intron structure of circRNAs. A *de novo* reconstruction of the inner circle structure using `FUCHS`*denovo* showed a gain of information of 15 %. Furthermore, `FUCHS`*denovo* identified alternative splicing in 8 - 10 % of circRNAs. Performing a differential motif enrichment analysis of the flanking introns of circRNAs with alternative splicing over circRNAs without alternative splicing identified FOXO as a potential transcription factor driving alternative splicing in circRNAs. Binding motifs for CPEB1 and HOX were enriched in the flanking introns of

circRNAs from host genes expressing many circRNAs over circRNAs from host genes expressing only one circRNA. To exemplify the value of the reconstructed circRNA models in downstream analyses, I performed a miRNA seed search and RBP motif search. Comparing the seed density of circRNAs and mRNAs showed that circRNAs were more densely populated with both, miRNA seeds and RBP motifs. This suggests that circRNAs could form an additional layer in the gene-regulatory network by competing with their host genes for miRNA or RBP binding.
Availability: https://github.com/dieterich-lab/FUCHS.git

Acknowledgements

First of all, I wish to thank my supervisors, Prof. Dr. Christoph Dieterich and Dr. Dario Valenzano, for their guidance during my time as Ph.D. student. I am grateful for their advice, ideas, and valuable discussions.
I would like to thank Prof. Dr. Andreas Beyer for his willingness to examine this work as well as Prof. Dr. Juliette de Meaux and Dr. Jorge Boucas for their willingness to join my defence committee.
Furthermore, I would like to thank the whole Dieterich Lab, especially Dr. Tobias Jakobi for always being there for me when I needed help and packaging `FUCHS` into an installable Python package. The whole Valenzano Lab for helpful discussions and support, especially Dr. Yumi Kim who taught me everything about molecular biology I wanted to know. The Bioinformatics Core Facility, especially Dr. Jorge Boucas, for their never ending computational support.
I would like to thank our collaborators, Prof. Dr. Gerhad Schratt and Dr. Marek Rajman, for a fruitful collaboration on synaptic plasticity in rat hippocampal neurons which enhanced my education.
I would like to acknowledge the DFG Priority Program SPP1738 "Emerging Roles of Non-coding RNAs in nervous system development, plasticity and disease" as well as the Max Planck Institute for Biology of Ageing for their financial support.
Special thanks goes to my parents and my grandma for always believing in me, for their constant support and everything they taught me. Many thanks to Melanie M., Ariadna P., Brittany G., and Dennis S. for always listening to me and always having my back. My volleyball team for the right balance between work and exercise. My extended family, especially my family in Minnesota for their support and the possibilities they facilitated.
Last but not least, I would like to thank everybody who helped me become who I am and where I am today. I am eternally grateful.

Table of contents

List of figures

List of tables

List of abbreviations

A3SS	Alternative 3' spice site
A5SS	Alternative 5' spice site
AEU	Alternative exon usage
AGO-PARCLIP	Argonaute PAR-CLIP
AS	Alternative splicing
BP	Base pair
BSJ	Back-splice junction
circExon	Circularized exon
circRNA	Circular RNA
CNS	Central nervous system
DNA	Deoxyribonucleic acid
EMT	Epithelial mesenchymal transition
ES	Exon skipping
FDR	False discovery rate
FP	False positive
GFP	Green fluorescent protein
GO	Gene ontology
IR	Intron retention
lncRNA	Long non-coding RNA
miRNA	Micro RNA
mRNA	Messenger RNA
PAR-CLIP	Photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation
PCR	Polymerase chain reaction

qPCR	Quantitative PCR
RCM	Reverse complementary matches
RNA	Ribonucleic acid
RNA-Seq	RNA-Sequencing
rRNA	Ribosomal RNA
siRNA	Small interfering RNA
snoRNA	Small nucleolar RNA
SNP	Small nucleotide polymorphism
TER	Transcription elongation rate
TF	Transcription factor
TP	True positive
tRNA	Transfer RNA
TSS	Transcription start site

Chapter 1

Introduction

If deoxyribonucleic acid (DNA) is the blueprint to build most living organisms, than ribonucleic acid (RNA) are partial copies of this blueprint which may be modified to fit the requirements of individual cells without altering the DNA. RNAs are single stranded molecules transcribed from the DNA by RNA polymerases. They consist of a ribose-phosphate backbone and four nitrogenous bases, adenine, cytosine, guanine, and uracil. These single strand molecules may form hydrogen bonds between adenine and uracil, cytosine and guanine, and guanine and uracil to form complex secondary and tertiary structures [1].

Each RNA copy of the DNA can be modified. There are over 110 types of RNA modifications that can affect the activity, localization and stability of the molecules. These modifications include methylation of a single base pair or an hydrolytic deamination changing adenine into inosine [2, 3].

The RNAs which are being translated into proteins are called messenger RNAs (mRNA). Besides this large class of coding RNAs there are several other classes of RNA performing different tasks. These RNAs are classified by their purpose, localization, length, or structure. RNAs forming a major component of the ribosome, the complex translating mRNA to proteins, are called ribosomal RNAs (rRNA). RNAs binding single amino acids to transport them to the ribosomes are called transfer RNAs (tRNA). Additionally, there are many regulatory RNAs such as very short RNAs called micro RNAs (miRNAs) or very long RNAs called long non-coding RNAs (lncRNA). There are small interfering RNAs (siRNA) playing a role in mRNA degradation [4]. There are small nucleolar RNAs (snoRNAs) guiding RNA modifications [5]. All of these RNAs are well studied classes of RNA. In my thesis I will focus on circular RNAs (circRNA), which are a unique class of RNA. Unlike all other linear

RNAs they form a covalently closed loop. Through improved sequencing strategies, they recently came back into the focus of molecular and computational biologists.

1.1 CircRNAs

Circular RNAs are a special class of RNA generally thought to be non-coding. In contrast to all other spliced RNA classes, circular RNAs arise when a 3' splice donor loops back to form a covalent bond with an upstream 5' splice acceptor instead of a linear downstream 5' splice acceptor (see Figure 1.1). Hence it is called a back-splicing event. Having neither a cap nor a poly-A tail, circular RNAs are more stable than linear RNAs. The median half-life of circular RNAs ranges from 18-24 hours which is a three fold increase compared to their linear equivalent (4-7.4 hours) [6]. Furthermore, circular RNAs are *RNAseR* resistant and are not present in poly-A enriched RNA libraries. The majority of circular RNAs are untranslated splice isoforms of protein coding genes which are called 'host genes'. Until today, every living organism that has been studied in the context of circular RNAs, from yeast and cell cultures to mice and monkeys, express circular RNAs [7].

Before I introduce my own work, I will give an introduction into the key features of circRNA history, biogenesis, function, and biomarker potential.

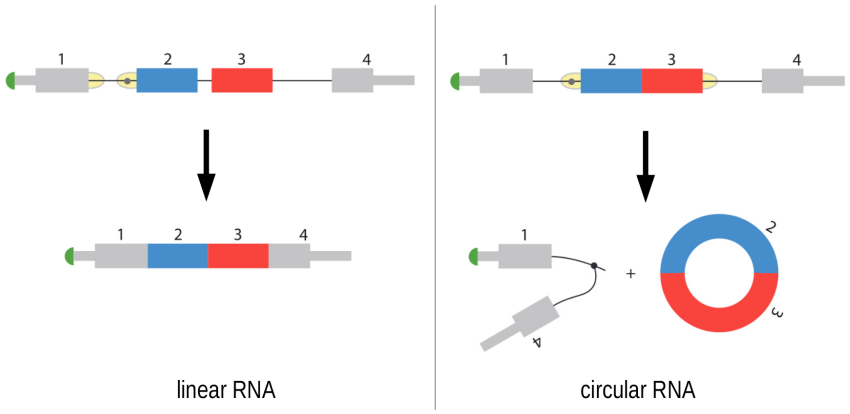

Figure 1.1: CircRNA formation. The left-hand side of the figure shows the formation of a linear mRNA with a 5' cap and a 3' poly-A tail. The right-hand side of the figure shows the formation of a circular RNA. It is a covalently closed RNA molecule without a 5' cap and poly-A tail. [modified from [8]]

1.2 The short history of circRNAs

The first report on circular RNA molecules dates back to 1979 when Coca-Prados *et al.* [9] observed circular RNA molecules in HeLa cells under an electron microscope (Figure 1.2). Their experiments showed that the majority of circRNAs localizes to the cytoplasm.

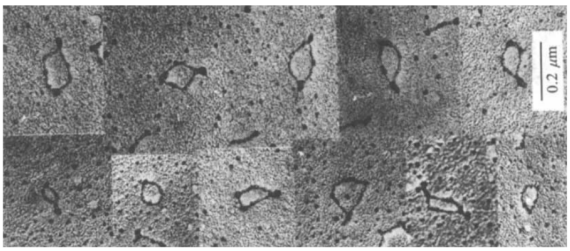

Figure 1.2: Electron microscopy image. This is the first picture of circular RNAs visualized through electron microscopy (modified from Coca-Prados *et al.* [9])

A decade later, Nigro *et al.* [10] noticed scrambled exons during an exhaustive study to further characterize the structure of the tumor suppressor gene DCC. They observed that not all transcripts followed a linear concatenation of exons but that a back splicing between a downstream 3' splice donor and an upstream 5' splice acceptor occurred at an estimated rate of 1 in every 1000 linear transcripts. They concluded that these transcripts were not trans-spliced because trans-splicing was only observed in lower eukaryotes.

Cocquerelle *et al.* [11] identified scrambled exons in the human ETS-1 gene when comparing the locus to the chicken c-ets-1 homologue. They performed PCR of both poly-A^+ and poly-A^- libraries and detected scrambled exon transcripts only in the poly-A^- fraction. They hypothesized that these transcripts were of no biological relevance but suggested that studying these splicing events might give further insights into the regulation of alternative splicing. Bailluel *et al.* [12] later revised this statement saying that a functional role of circular RNAs could not be ruled out.

In 1993, Capel *et al.* [13] proved that the sex determining gene Sry expresses circular transcripts and for the first time showed tissue-specific expression of circular RNAs. They confirmed that the circular form of Sry located to the cytoplasm and was the dominant form over the linear form in adult testis, whereas the linear form was more abundant during development. They concluded that the formation of circular junctions could regulate the translation of Sry in adult testis.

After 1993, other groups reported the presence of single genes capable of expressing scrambled exons soon termed exon circles and finally circular RNA (circRNA).
Until the beginning of 2010, only a few groups explored the circular RNA landscape in different organisms, but with recent changes and advances in standard transcriptomic analysis our knowledge about circRNAs is expanding quickly.

1.3 Circular RNAs are regulated through different mechanisms

Ashwal-Fluss *et al.* [14] and Starke *et al.* [15] showed that circularization happens co-transcriptionally and depends on the canonical splicing machinery, canonical splice sites, and intron sequences. They concluded this from knockdown experiments, mutation of canonical splice sites, and mutation of intron sequences.
Liang *et al.* [16] extended this concept by creating several ZKSCAN1 transcripts with variable lengths of the flanking *Alu* repeats. Thus, they determined the minimal *Alu* sequence necessary in the flanking introns to drive circularization. Furthermore, they cloned other exons in between these minimal *Alu* sequences and validated the circularization of these exons. This research resulted in a commercially available plasmid with the minimal *Alu* repeats and restriction enzyme sites to allow for the integration of custom exons to be circularized. This has been tested in HeLa and Hu7 cells. It is yet unclear if this vector generates circRNAs in other organisms or *in vivo*.
Apart from the general observation that reverse complementary matches (RCMs) in flanking introns drive circularization, several RNA binding proteins (RBP) were indicated to influence circularization. For example, the splicing factor muscleblind (MBNL1) correlates with the expression of circMBL [17, 14]. In contrast to that, Conn *et al.* [18] and Ivanov *et al.* [19] pursued a more large scale approach to identify factors associated with circularization.
Conn *et al.* [18] investigated the circRNA biogenesis during epithelial mesenchymal transition (EMT) and identified Quaking (QKI) as the main driver for exon circularization using their circScreen approach. CircScreen is a SMARCA5 gene construct modified to express *GFP* if exon 16 splices back to exon 15 or *mCherry* if exon 16 splices forward to exon 17. They transfected HEK293T cells and induced EMT via two siRNAs in independent experiments. They measured the nuclear fluorescence

of 20 possible factors such as APOBEC3B, MBNL1, and NOVA1 but only QKI showed an effect on the *GFP* levels. Strikingly, QKI itself was regulated during EMT. They found QKI binding motifs in the flanking introns of circularized exons. Through dimerization of QKI, splice sites come into close proximity so that the splice machinery is likely to circularize exons flanked by QKI motifs (Figure 1.3 a). The same was observed when their circScreen reporter was cloned into four other genes validating their observation that QKI drives circularization. However, the absence of an effect of MBNL1 does not contradict previous observations by Ashwal-Fluss *et al.* [14]. It rather underlines the complexity as well as potential specificity of circRNA expression.

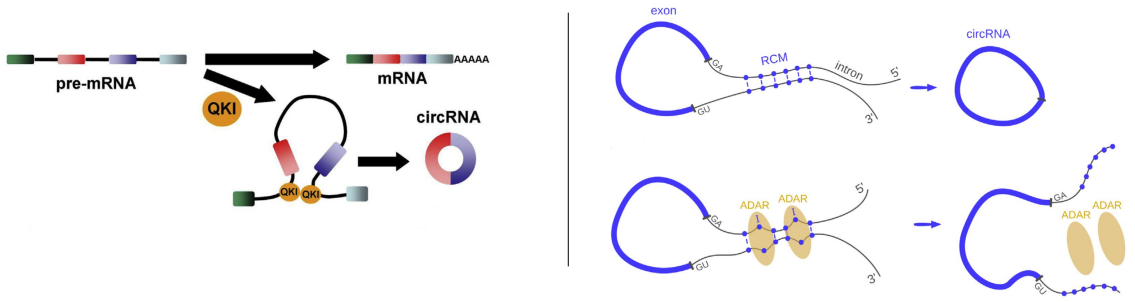

Figure 1.3: Biogenesis of circRNA. The left-hand side of the figure shows how QKI binds to the flanking introns. Through dimerization of QKI the splice sites are brought into close proximity and a circular RNA is formed [18]. The right-hand side of the figure shows how RCMs cause the flanking introns to bind to each other, bringing the splice sites into close proximity and a forming a circular RNA. ADAR causes double stranded RNA editing, thus melting the hairpin structure. No circular RNA is formed [19].

Ivanov *et al.* [19] pursued a more computational, genome-wide approach. Blasting intron sequences of C. *elegans* against each other revealed that introns flanking circularized exons are significantly enriched for RCMs compared to random introns of equivalent length. Based on this knowledge, Ivanov *et al.* designed a circularization score (H) and used this score to predict the potential of introns to form circular RNAs. The predictions were validated on six high-scoring, ten medium-scoring, and five low-scoring genes. Twelve out of 16 high- and medium-scoring genes yielded circRNAs. To prove that RCMs were driving circularization, they knocked down the double stranded RNA-editing protein ADAR, as well as the single strand RNA-editing protein APOBEC, hypothesizing that ADAR would melt the hairpin structures formed by the RCMs using APOBEC as negative control (Figure 1.3 b). Using siRNAs, they observed a stronger increase of selected circRNAs upon ADAR knockdown compared to APOBEC and mock control.

A more recent study by Zhang *et al.* [20] showed that the speed of Polymerase II (Pol II) is correlated with circRNA expression. This study shows that the transcription elongation rate (TER) of genes yielding circRNAs is significantly higher than in non-circularized genes. Thus, it is hypothesized that a fast elongation rate favors circRNA formation by allowing flanking introns to form hairpin structures through the aforementioned mechanisms (Quaking, MBNL, RCMs). To test this hypothesis, the relative expression of selected circRNAs was measured in E1126G and R749H mutants, causing increased TER and decreased TER respectively. Indeed, faster Pol II led to an increase in relative circRNA expression. This increase was not observed in the slower Pol II mutant.
In summary, circRNA biogenesis depends on various factors highlighting the potential to be specifically regulated in different tissues and during different processes. Most strikingly, these different mechanisms have one element in common; they all show that the splice donor and acceptor are brought into close proximity to form a back-splice junction.

1.4 Functional circRNA studies are scarce

Although the existence of circRNAs has been known for more than three decades, we are still at the beginning of circRNA research, thus not much is known about their function. Only for a few well-researched circRNAs such as CDR1-AS and MBNL a function has been described. In a study on the regulation of cerebellar degeneration-related protein 1 (CDR1) by miR-671 CDR1-AS was identified as a circular anti-sense transcript of CDR1 [21]. Hansen *et al.* showed that CDR1-AS is a direct target of miR-671 leading to the cleavage of the circle structure. Furthermore, they showed a positive correlation of the expression levels of CDR1 and CDR1-AS. Two years later, two independent studies revealed that CDR1-AS harbors over 70 miR-7 seeds, thus concluding that CDR1-AS has regulatory roles acting as a miR-7 sponge [22, 23]. After showing that CDR1-AS and miR-7 co-localize in the cytoplasm, Hansen *et al.* [21] measured the response of known miR-7 targets, SCNA, EGFR, and IRS2, to miR-7 expression with CDR1-AS and without CDR1-AS. Indeed, in the presence of the circRNA miR-7 targets responded less efficiently than in cell lines without CDR1-AS.
Memczak *et al.* [23] used a different approach to show the sponging potential of CDR1-AS. They performed an AGO-PARCLIP [24, 25], and observed a strong signal of CDR1 anti sense reads, while

PARCLIP control RBPs yielded virtually no CDR1 anti sense reads. Furthermore, they showed that a miR-7 knockdown had a similar phenotype in zebrafish brain as CDR1-AS expression in zebrafish brain, a transcript absent in wild type zebrafish. These findings demonstrated that CDR1-AS acts as a miR-7 sponge. The current hypothesis is that CDR1-AS binds miR-7. Then the complex localizes to the cytoplasm where, upon a miR-671 induced cleavage of the circRNA, the miRNAs are released. Till today this is the only known circular miRNA sponge. However, this indicates that there could be other circRNAs sponging miRNAs.

The circulatization of the second and third exon of the splicing factor muscleblind (MBL) was first discovered by Housely *et al.* in 2006 [17] but only a decade later Ashwal-Fluss *et al.* [14] showed that circMBL harbors binding sites for the RBP MBL/MBNL1. The binding of MBL to circMBL is potentially responsible for gene regulation by competition between circular and linear splicing.

The lack of functional experiments for circRNAs shows that the field of circRNA research is still at its beginning.

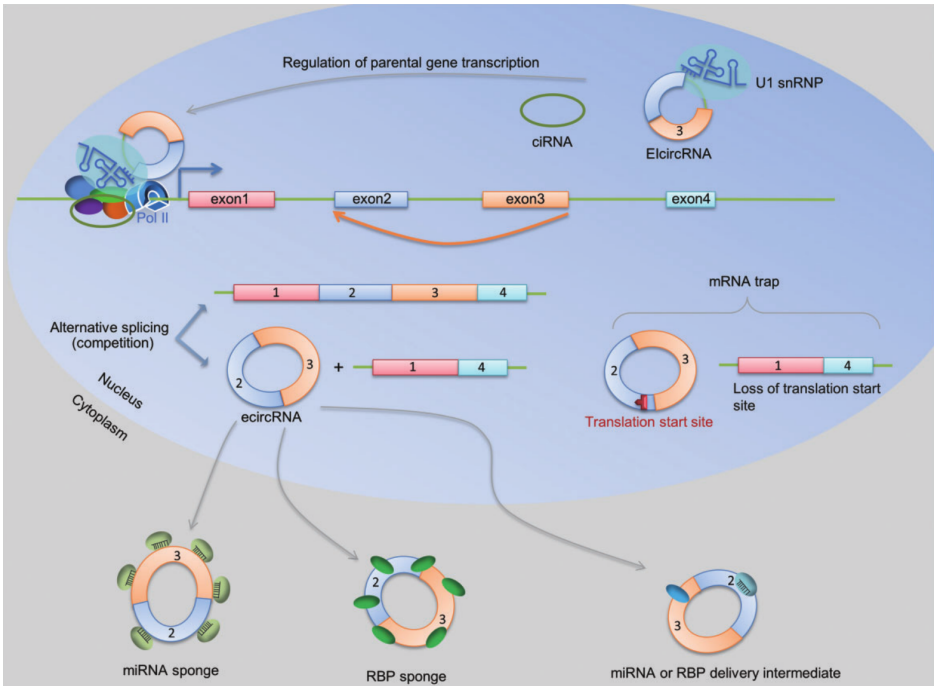

Figure 1.4: Known functions of circular RNA. This figure shows an overview of the currently known functions of circRNAs. There are circRNAs in the cytoplasm possibly acting as miRNA and RBP sponge or shuttle platform. CircRNAs in the neucleus can contain a intron sequences which have been indicated to regulate the transcription of their host genes. An mRNA trap arises if the circularized exons contain the translation start site (TSS), thus the mRNA loses its TSS or may need to use an alternative TSS if possible. [26]

1.5 CircRNAs are linked to cancer and other diseases

Recently, many cancer studies focused on the expression of circular RNAs in cancerous versus normal cells. Though they do not provide any functional or mechanistic insights to how circRNAs are involved in cancer formation, these studies indicate that circRNAs may be used as biomarkers. Wang *et al.* [27] showed a significant correlation of hsa_circ_001988 with colorectal cancer in 31 matched tumor/normal pairs. Predicting colorectal cancer based on hsa_circ_001988 levels achieved a specificity of 68 % and a sensitivity of 73 %. In a similar study Li *et al.* [28] predicted gastric cancer based on hsa_circ_002059 levels with a specificity of 62 % and a sensitivity of 81 %. In both studies the effect size was small (see Figure 1.5 a and b), raising the question of how reliable it is to use circRNAs as biomarkers.

A genome-wide association study of small nucleotide polymorphisms (SNPs) for atherosclerosis risk revealed polymorphisms on chromosome 9p21.3 [29]. One of the four identified causal variants in this locus leads to an increase in circular ANRIL, a linear long non-coding RNA shown to regulate the tumor suppressor gene INK4/ARF. The authors hypothesized that ANRIL loses its regulatory capacity when it is circularized, thus increasing the risk for atherosclerotic vascular disease.

A small study on miR-7 sponge CDR1-AS in patients with sporadic Alzheimer's disease showed that CDR1-AS is significantly downregulated in patients compared to healthy individuals [30] (Figure 1.5c). The authours hypothesized that the absence of CDR1-AS leads to an increase of free miR-7 which downregulates known Alzheimer's disease associated genes such as UBE2A.

All these studies, whether they indicate circRNAs as risk factors or as biomarkers, show the number and the complexity of processes circRNAs are likely to be involved in.

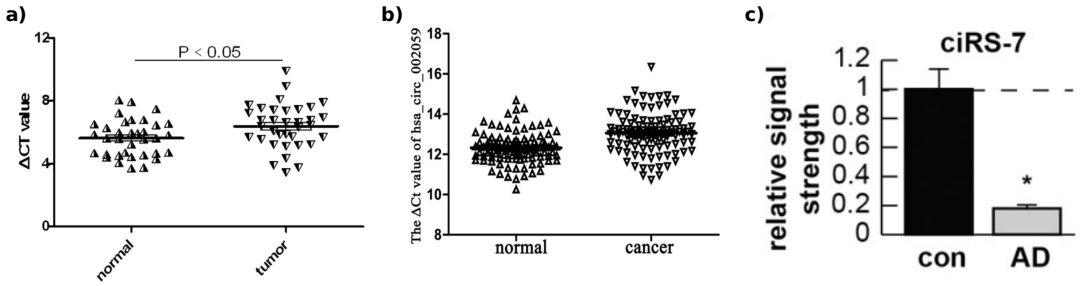

Figure 1.5: CircRNA expression as biomarker. a) hsa_circ_001988 in colorectal cancer [27]. **b)** hsa_circ_002059 in gastric cancer [28]. **c)** CDR1-AS in Alzheimer's disease [30].

1.6 My contribution

In the second chapter of my thesis, I will discuss how circRNAs are detected using RNA-Seq comparing three circRNA detection programs, including DCC which was developed in the Dieterich lab by Jun Cheng [31] and which is maintained by Dr. Tobias Jakobi. Subsequently, I will use DCC to detect circRNAs in murine heart and liver samples. In the third chapter, I will focus on my own program FUCHS [32], a pipeline to further characterize the inner structure of circular RNAs. I will use this pipeline to describe the circRNAs found in the mouse samples. In the fourth chapter, I will introduce the additional module FUCHS*denovo* which reconstructs the exon-intron structure based on intron signals and annotation. Again, I will demonstrate the use of the *de novo* chain reconstruction on the murine circRNAs. In the fifth chapter, I will use the output of my program to show how this information could be used to perform sequence based miRNA seed analysis or functional enrichment analysis. After summarizing my work on circular RNAs, I will finish my thesis by highlighting my work on the potential of the African turquoise killifish to study circRNAs in ageing.

In summary, my work presents a new pipeline to further characterize the inner structure of circular RNAs, a comprehensive analysis of the circular RNA landscape in murine heart and liver, as well as an initial analysis of the circular RNA landscape in ageing African turquoise killifish.

Chapter 2

CircRNA detection using RNA-Seq

2.1 Introduction

As next generation sequencing techniques became cheaper, scientists started to sequence whole transcriptomes and not only mature mRNA (rRNA depletion and poly-A enrichment). Sequencing the whole transcriptome revealed many chimerically spliced reads, which were initially filtered out as misaligned reads. A thorough analysis of these chimerically mapped reads led to the discovery of a plethora of genes forming circular RNAs [33]. Since then, developers of mapping programs recognized the abundance of circular RNAs and improved their algorithms to enhance the alignment of chimeric reads. This opened the door to the systematic investigation of circRNAs in different organisms. The first publications on transcriptome-wide circRNA expression used custom scripts to filter and quantify chimeric reads. These custom scripts were based on identifying back-spliced reads as depicted in Figure 2.1 [8, 34].

The need for standardized circRNA detection programs grew rapidly. In 2015 and early 2016, several circRNA detection programs were published (see Table 2.1). The main difference between these detection programs is that they facilitate the annotation of chimeric reads written by different mapping programs as there is no standard representation of chimeric reads yet.

In the following chapter, I will first compare three of the most recent circRNA detection programs (see Table 2.1 `CIRI`, `KNIFE`, and `DCC`) and then use the best detection tool to detect circular RNA in a new mouse dataset.

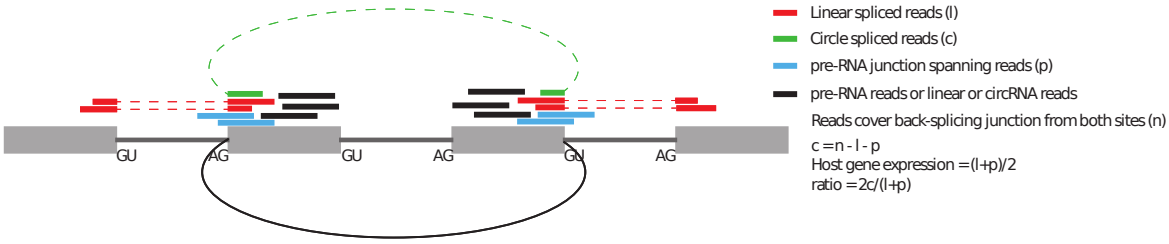

Figure 2.1: Read assignment and counting. Reads spanning the back-splice junction (c) are shown in green. The coverage of a host gene is measured by averageing the number of reads spanning the flanking introns; linear-spliced reads (l) and exon-intron reads (p) are summed and averaged over both flanking introns. (Adapted from [31])

Table 2.1: State of the art circRNA detection tools

Tool	Function	Mapper	Year	Reference
`find_circ`	circRNA detection	`Bowtie` [35]	2013	Memczack *et al.* [23]
`segemehl`	circRNA detection	`segemehl` [36]	2014	Hoffmann *et al.* [36]
`CIRCexplorer`	circRNA detection	`Bowtie` [35]	2014	Zhang *et al.* [37]
`circRNA_finder`	circRNA detection	`STAR` [38]	2014	Westholm *et al.* [34]
`CIRI`	circRNA detection	`BWA` [39]	2015	Gao *et al.* [40]
`KNIFE`	circRNA detection	`Bowtie` [35]	2015	Szabo *et al.* [41]
`DCC/CircTest`	circRNA detection and differential expression	`STAR` [38]	2016	Cheng *et al.* [31]

2.2 Benchmarking circRNA detection tools

`DCC` was developed in the Dieterich Lab by Jun Cheng. Though I was not involved in the development of the program, I performed the benchmarking of the program. I compared `DCC` to `CIRI` and `KNIFE` on two datasets. This analysis was published as a supplement of the `DCC` publication [31].

2.2.1 Data

The first dataset used for benchmarking was a mouse dataset [31]. A matched $RNaseR^{+}$ and $RNaseR^{-}$ sample pair was sequenced with 2 x 100 BP long reads on an Illumina HiSeq2500 (see Table 2.2). The reads were mapped to the mouse genome assembly GRCm38 using the ENSEMBL release 79 genome annotation.

The second dataset was a HeLa cell culture dataset that was released with the `CIRI` publication [40].

Two technical replicates of each *RNaseR*$^+$ and *RNaseR*$^-$ were sequenced with 2 x 100 BP long reads on an Illumina HiSeq2000 (see Table 2.2).[1]

Table 2.2: Benchmarking data

Species	Library-prep	Experiment	Platform	Acccession
BALB/c mouse	ScriptSeq v2 Epicentre	*RNaseR*$^-$	Illumina HiSeq2500	SRX1165561
BALB/c mouse	ScriptSeq v2 Epicentre	*RNaseR*$^+$	Illumina HiSeq2500	SRX1175091
HeLa	Illumina TrueSeq	*RNaseR*$^-$	Illumina HiSeq2000	SRR1637089
HeLa	Illumina TrueSeq	*RNaseR*$^-$	Illumina HiSeq2000	SRR1637090
HeLa	Illumina TrueSeq	*RNaseR*$^+$	Illumina HiSeq2000	SRR1637085
HeLa	Illumina TrueSeq	*RNaseR*$^+$	Illumina HiSeq2000	SRR1637086

2.2.2 Methods

CircRNAs were first detected in *RNaseR*$^-$ samples and then validated in matched *RNaseR*$^+$ samples. Comparing matched samples created an advantage over comparing independent samples; it reduced errors due to biological variance. If a circle was expressed in one tissue, it would be present in both libraries of matched samples, however not in two independent samples. The CircleSeq protocol [33] provided a gold standard. It involves one step in which one rRNA depleted library is split into two libraries but only one fraction is further treated with *RNaseR*. Thus, false positive circRNAs were circRNAs which were present only in the *RNaseR*$^-$ and not the *RNaseR*$^+$ sample.

Program calls

In the following section I will list all command line calls that were executed for this benchmarking. All samples were subjects of the same sequence of program calls.

`DCC`

To detect circRNAs using `DCC`, `flexbar` [42] was used to trim the adapters and poor quality bases (Phred score ≤ 20). Trimmed reads were aligned using `STAR` [38]. For paired-end data, the reads of each mate were mapped independently and jointly in one run. `DCC` uses selected `STAR` output files to detect back-splice junctions, i.e. Chimeric.junction.out, SJ.out.tab, Aligned.sortedByCoord.out.bam.

[1] Sample information was confirmed through personal communication with Yuan Gao, first author of `CIRI`.

Additionally, it requires a reference sequence. Providing annotation files allowed DCC to simultaneously filter and annotate circRNA candidates. Since DCC reports the strand of circRNAs, it is important to declare the library's strandedness with the -ss/-N flag, indicating that the data is second strand, or unstranded respectively. This flag was only used for the mouse data. CircRNAs from HeLa data were detected without setting either flag as they were sequenced with a first strand RNA-Seq protocol. A detailed list of parameters for `flexbar`, `STAR`, and `DCC` is outlined below.

```
# Trimming
flexbar -n 4 -r [Sample]_R1.fastq.gz -p [Sample]_R2.fastq.gz
        -t [Sample] -f sanger -u 5 -m 20 -z GZ -q 20 -as AGATCGGAAGAGC

# Mapping
STAR  --runThreadN 10 --genomeDir [Genome]
      --outSAMtype BAM SortedByCoordinate
      --readFilesIn [Sample]_R1.fastq.gz [Sample]_R2.fastq.gz
      --readFilesCommand zcat --outFileNamePrefix [Sample]
      --outSJfilterOverhangMin 15 15 15 15 --alignSJoverhangMin 15
      --alignSJDBoverhangMin 15 --seedSearchStartLmax 70
      --outFilterMultimapNmax 2 --outFilterScoreMin 1
      --outFilterMatchNmin 1 --outFilterMismatchNmax 2
      --chimSegmentMin 15 --chimScoreMin 15
      --chimScoreSeparation 10 --chimJunctionOverhangMin 15
      --alignTranscriptsPerReadNmax 50000

# CircRNA detection
DCC @sample_sheet -mt1 @mate1 -mt2 @mate2 -D -an [Annotation].gtf
    -R [Repeats].gtf -M -Nr 2 1 -fg -temp -ss -F -Pi -L 20
```

CIRI

The version of CIRI (v1.2) used for this benchmarking required all reads to have the same length. Hence, flexbar was used with -m 100 and -k 100 to guarantee that all reads were 100 BP long. Reads were mapped using BWA [39] with the parameters suggested in the CIRI tutorial, which were optimized for circle detection using CIRI. CIRI uses SAM files produced by BWA as well as a reference FASTA and an annotation in GTF format. The -U 3 option was used for CIRI to discard all multi mapping reads. A detailed list of parameters for flexbar, bwa mem, and CIRI is outlined below.

```
# Trimming
flexbar -n 4 -r [Sample]_R1.fastq.gz -p [Sample]_R2.fastq.gz
        -t [Sample] -f sanger -u 5 -m 100 -k 100 -z GZ

# Mapping
bwa mem -P -T 19 -t 18 -B 3 [Genome.index] [Sample]_R1.fastq.gz
        [Sample]_R2.fastq.gz > [Sample].sam

# CircRNA detection
perl CIRI_v1.2.pl -U 3 -I [Sample].sam -O [Sample].minusU.ciri
                  -F [Genome].fa -A [Annotation].gtf -P -high
```

KNIFE

The KNIFE manual recommended trimming of the adapters and removing poor quality bases. Again, flexbar was used with the same parameters as for DCC. KNIFE was installed for SLURM [43] and the pipeline ran with the suggested parameters setting the splice junction overhang to 13 and the mode to sam. The user was required to run KNIFE twice to detect and quantify all back-splice junctions. findCircularRNA.sh started the mapping with Bowtie [35] as well as the circle detection. A script to perform an additional filtering step to filter for circRNAs with p_predicted ≥ 0.9 (posterior probability) was implemented and candidates from

the [sample]_report.txt and unaligned_[sample]_denovo_report.txt were merged. A detailed list of parameters for `flexbar` and `KNIFE` is outlined below.

```
# Trimming
flexbar -n 4 -r [Sample]_R1.fastq.gz -p [Sample]_R2.fastq.gz
        -t [Sample] -f sanger -u 5 -m 20 -z GZ -q 20 -as AGATCGGAAGAGC

# KNIFE first run
sh findCircularRNA.sh [Read/directory] [Output/folder]
                      [Sample] 13 sam_[organism]_large

# KNIFE second run
sh findCircularRNA.sh [Read/directory] [Output/folder]
                      [Sample] 13 sam_[organism]_large_unaligned
```

Evaluation parameters

I evaluated all programs' performances based on runtime, precision, agreement, and quantification agreement.

Runtime

The runtime was estimated on the mouse dataset as proof of principle. All programs tested in this benchmark used different read aligner: `DCC` used `STAR`, `CIRI` used `BWA` and `KNIFE` used `Bowtie2`. This is why the runtime was measured on both, the read alignment and circRNA detection step. To better understand the difference in runtime, Figure 2.2 stratifies the runtime by process. All programs were executed in identical compute environments (Intel Xeon CPU E5-4640 0 @ 2.40GHz, 256 GB RAM, 64 cores).

Precision of circRNA detection

To estimate the precision of each tool I calculated the false discovery rate by detecting circRNAs in *RNaseR*$^-$ samples and then validating these candidates in the *RNaseR*$^+$ treated samples. CircRNAs that could not be validated with the *RNaseR*$^+$ samples were classified as false positives.
For `STAR/DCC` and `BWA/CIRI` the precision was also evaluated on single-end versus paired-end sequencing strategies. This was done by performing an analysis where the paired reads of the mouse data were uncoupled to simulate single-end libraries. I compared these single-end runs to the standard paired-end runs.

Agreement

Besides the precision, the agreement between the programs on the circRNA candidates was evaluated as well. The agreement was measured on the reported back-splice junction coordinates as well as the expression level of each commonly identified back-splice junction.

2.2.3 Results and discussion

`STAR/DCC` is the fastest circRNA detection program

Comparing the runtime, the combination of `DCC` and `STAR` emerged as the fastest circle detection pipeline. The combination of `STAR` and `DCC` ran fastest in both cases (*RNaseR*$^{+/-}$; Figure 2.2), while `CIRI` was the slowest circle detection program even though `BWA` performed similarly to `STAR`. `CIRI`'s runtime was strongly affected by the number of candidate circles (*RNaseR*$^+$: $\sim$ 10,000 circRNAs; *RNaseR*$^-$: $\sim$ 2,500 circRNAs) and `KNIFE`'s runtime depended most on the number of input reads (*RNaseR*$^+$: 19,960,612 reads; *RNaseR*$^-$: 29,157,170 reads). `DCC` benefited from `STAR`'s short runtime and made the most effective use of the alignment method's output. Thus, it was the fastest method regardless of the number of input reads, or the number of candidate circles.

`STAR/DCC` is the most precise circRNA detection program

Evaluating the precision of `DCC`, `CIRI`, and `KNIFE` on the mouse dataset showed that `DCC` was the most precise method with an FDR of 3 %. `DCC` also detected the highest number of true positives. `CIRI`

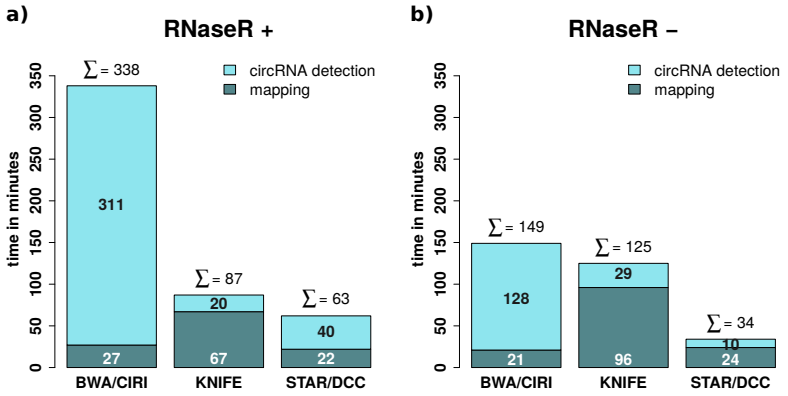

Figure 2.2: Runtime of benchmarkin the mouse dataset. a) shows the runtime on the *RNaseR*$^+$ sample with more circRNAs than the *RNaseR*$^-$ sample. `CIRI` performed exceptionally long compared to `KNIFE` and `DCC`. **b)** shows the runtime on the *RNaseR*$^-$ sample with more reads than the *RNaseR*$^+$ sample. For this sample `DCC` was approximately four times as fast as `CIRI` and `KNIFE`

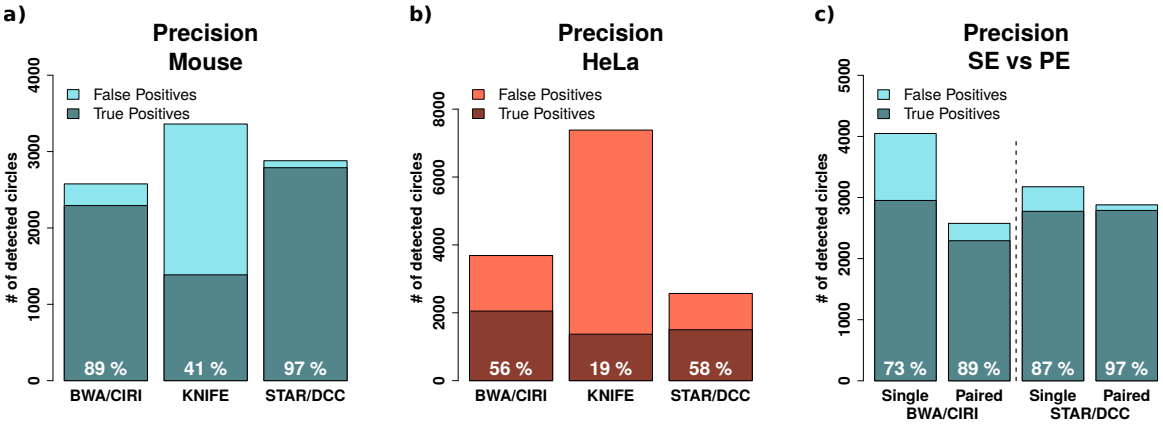

Figure 2.3: Precision a) Precision of benchmarking the mouse dataset. `DCC` is the most precise detection tool with an FDR of 3 % **b)** Precision of benchmarking the HeLa dataset. `DCC` is the most precise detection tool again. The large amount of false positives are likely to result from varying library sizes (SRR1636985: 13.3 mio reads; SRR1636986: 23.5 mio reads; SRR1637089: 45 mio reads; SRR1637090 35.7 mio reads). **c)** Comparing the precision of running the samples as paired-end or single-end shows that `DCC` is more precise than `CIRI` and that considering the paired information is superior to single-end reads.

performed reasonably well and the previously reported precision was confirmed [40]. `KNIFE` had the lowest precision of them all.

All three methods performed worse on the HeLa dataset compared to the mouse dataset. While `DCC` and `CIRI` were similarly precise, `KNIFE` identified 6013 circRNA candidates, which could not be validated in the matched *RNaseR* treated sample. The most likely reason for the poor performance was not the detection within the *RNaseR*$^-$ set but the varying sequence depth (SRR1636985: 13.3 mio reads; SRR1636986: 23.5 mio reads; SRR1637089: 45 mio reads; SRR1637090 35.7 mio reads).

Thus, the number of circRNAs was influenced by the sequencing depth and library protocol such that the precision of all programs could have been underestimated. Nevertheless, DCC is the most precise, even though it detects fewer circRNAs than CIRI within this dataset.
Comparing paired-end vs. single-end runs demonstrated that paired-end data had a higher precision while detecting similar amounts of true positive circRNAs. When using DCC the difference between single-end and paired-end data was negligible but when using CIRI the precision decreased by 16 % when omitting the mate information.

On average, BWA/CIRI attribute more reads to back-splice junctions than other programs

Figure 2.4 shows that all methods agree on a large set of true positive circRNAs with DCC and CIRI sharing the largest amount of true positive circRNAs. Furthermore, to investigate DCC's quantification of circRNAs compared to the other programs, Figure 2.5 contrasts the reads per junction detected with program *X* and *Y* for each pairwise comparison. On average, STAR/DCC assigned fewer reads to a back-splice junction than BWA/CIRI (Figure 2.5a and d). Comparing STAR/DCC to Bowtie2/KNIFE did not reveal an evident bias (Figure 2.5b and e).

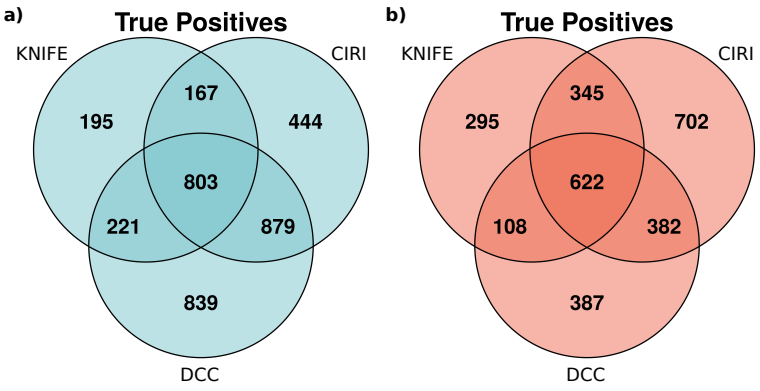

Figure 2.4: Agreement of the methods. Comparing only the true positive circRNAs shows that the programs do not agree well on the identified circRNAs. **a)** Mouse dataset **b)** HeLa dataset

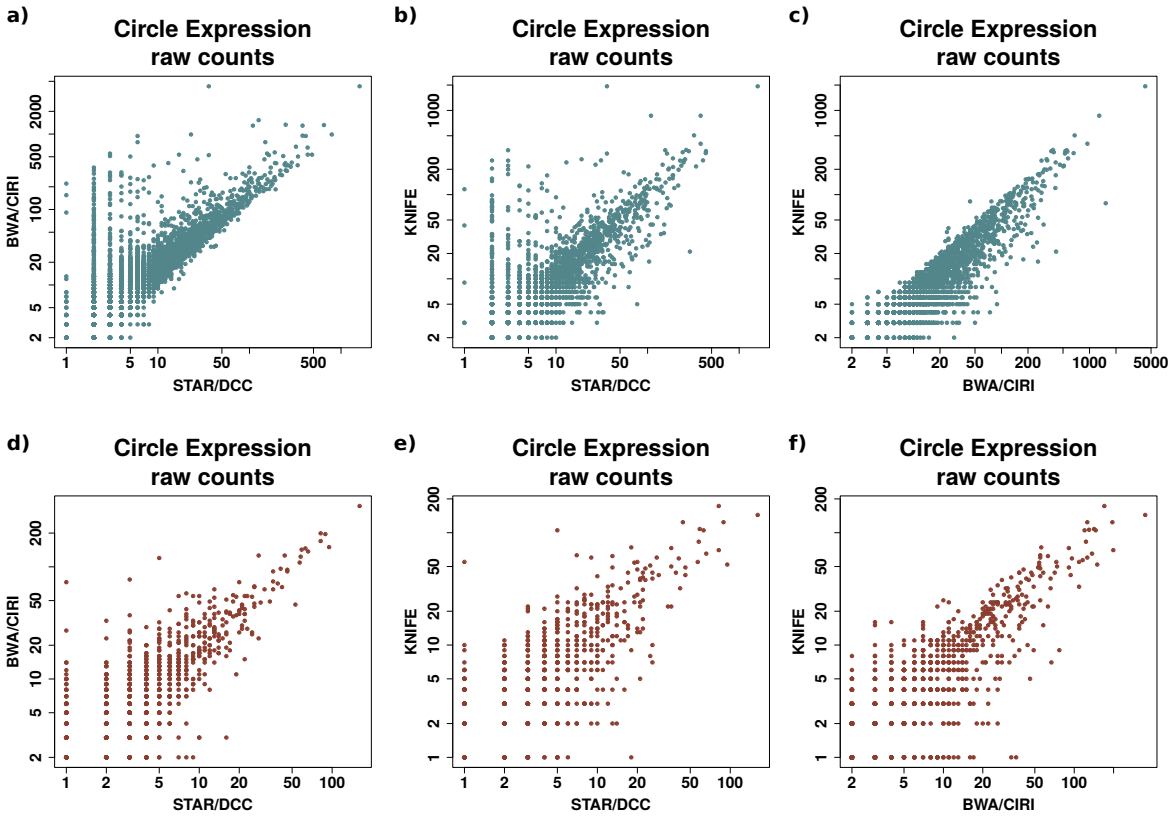

Figure 2.5: Comparison of junction quantification. Comparing the number of reads assigned to a junction by different programs. `BWA` (`CIRI`) assignes more reads to a junction than the other programs to the same junctions. **a-c)** Mouse dataset **d-f)** HeLa dataset

2.2.4 Conclusions

Based on the parameters I used to evaluate the performance and quality of the programs it became evident that `DCC` outperformed both `CIRI` and `KNIFE` in runtime and precision. While `DCC` and `CIRI` performed similarly well, it is questionable if `KNIFE` should be used for circRNA detection at all. The main difference between `DCC` and `CIRI` was that they used different mapping algorithms to detect circRNAs. `CIRI` is best to use when searching for all possible circRNAs accepting that 10-40 % of these will likely be false positives. `DCC` should be used when searching for highly confident circRNAs accepting that some circRNAs will be missed.

2.3 Comprehensive circRNA study on long reads

Due to the low abundance of circular RNA compared to linear RNA, a large amount of circular RNAs is potentially overlooked when detecting circRNAs from rRNA depleted samples. Hence, for the remainder of my thesis I will be working on rRNA depleted, *RNaseR* treated libraries from mouse. First, I will describe the general circRNA landscape in these samples and then I will use this dataset to demonstrate the usage of the pipeline I developed.

2.3.1 Data

To discover the full spectrum of the circular RNA diversity in murine heart and liver, two heart and two liver samples were enriched for circular RNA molecules by *RNaseR* treatment. They were sequenced with paired-end, 250 BP long reads on an Illumina HiSeq2500 to increase the chance of sequencing back-splice junctions. In the following section, I will highlight the results of detecting circular RNA in these samples using `DCC`, as it was shown to have the best performance during the benchmarking. Table 2.3 provides an overview of the samples used throughout the rest of the thesis.

Table 2.3: Long reads mouse data

Sample	Tissue	Age	Strain	Experiment	Accession
Y_HE	Heart	10 months	C57BL/6	*RNaseR*$^+$	SRX2504989
O_HE	Heart	27 months	C57BL/6	*RNaseR*$^+$	SRX2504987
Y_LI	Liver	10 months	C57BL/6	*RNaseR*$^+$	SRX2504990
O_LI	Liver	27 months	C57BL/6	*RNaseR*$^+$	SRX2504988

2.3.2 Methods

Read Trimming: `flexbar`

Only `STARlong` is able to map reads longer than 249 base pairs for one end. However, `STARlong` is not capable of mapping chimeric reads. Hence, the reads were mapped using `STARshort`. Therefore, all reads had to be trimmed to 249 base pairs using `flexbar` version 2.5 with the following parameters.

```
flexbar -n 4 -r sample_R1.fastq.gz -p sample_R2.fastq.gz
        -t sample -f sanger -u 50 -k 249 -z GZ
```

Read Mapping: `STAR`

All RNA-Seq reads were mapped using `STARshort` version 2.5.1b to the mouse genome, GRCm38 (mm10), using the Ensembl release 79 reference annotation. As mentioned before, `DCC` requires paired-end reads to be mapped as paired-end reads and as single-end reads. In this case, only the unmapped reads were mapped as single-end reads. All three runs were performed using the same parameters.

```
STAR --readFilesCommand zcat --runThreadN 18
     --genomeDir [genome] --outSAMtype BAM SortedByCoordinate
     --readFilesIn [sample]_1.fastq.gz ([sample]_2.fastq.gz)
     --outFileNamePrefix [sample] --quantMode GeneCounts
     --genomeLoad NoSharedMemory --outReadsUnmapped Fastx
     --outSJfilterOverhangMin 15 15 15 15 --alignSJoverhangMin 15
     --alignSJDBoverhangMin 10 --outFilterMultimapNmax 20
     --outFilterScoreMin 1 --outFilterMismatchNmax 999
     --outFilterMismatchNoverLmax 0.05 --outFilterMatchNminOverLread 0.7
     --alignIntronMin 20 --alignIntronMax 1000000
     --alignMatesGapMax 1000000 --chimSegmentMin 15 --chimScoreMin 15
     --chimScoreSeparation 10 --chimJunctionOverhangMin 15
     --twopassMode Basic --alignSoftClipAtReferenceEnds No
     --outSAMattributes NH HI AS nM NM MD jM jI XS
     --sjdbGTFfile [annotation].gtf
```

CircRNA Detection: `DCC`

CircRNA candidates were detected using `DCC` version 0.4.4 with the Ensembl annotation release 79. CircRNAs were detected with a threshold of at least two reads in at least two samples.

```
DCC @samplesheet -T 18 -ss -D -an GRCm38.79.clean.gtf -Pi
    -mt1 @mate1 -mt2 @mate2 -F -M -R GRCm38.79.repeats.clean.gtf
    -Nr 2 2 -fg -G -B @samplebam -A GRCm38.dna.toplevel.fa
```

Functional annotation: `GOrilla`

To obtain a first impression in which processes circRNAs could be involved in, I performed a gene ontology (GO) enrichment analysis using `GOrilla` [44]. Essentially, `GOrilla` first assigns the relevant GO terms to each gene. Next, it calculates an enrichment score for each GO term to compare the proportion of target genes in the GO term over the proportion of background genes in the GO term as visualized in Figure 2.6. Then, *p*-values are calculated based on a hyper geometric model and are corrected for multiple-hypothesis testing using the Benjamini-Hochberg method [44].

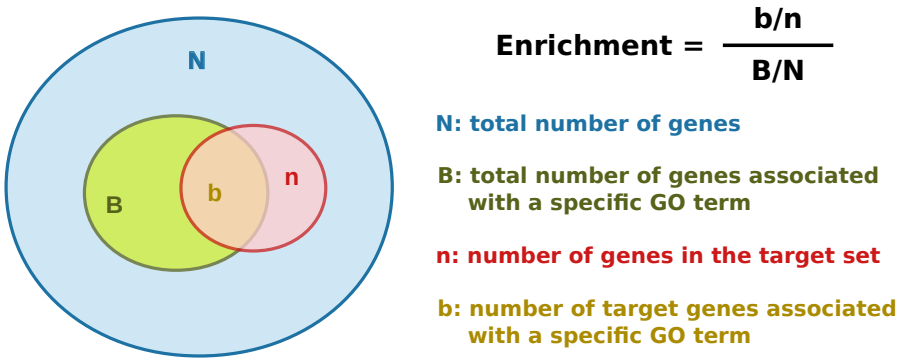

Figure 2.6: GO enrichment score E. Schematic view of how the GO enrichment score is calculated.

Genes expressing circRNAs in both samples of either heart or liver tissue were compared against all genes that were expressed in at least one sample of the respective tissue. Furthermore, I compared host genes of circRNAs expressed in all samples against genes that were expressed in at least one sample.

2.3.3 Results and discussion

RNaseR treatment digests linear RNA from their free 3' end. Since circular RNAs do not have a free 3' end the exo-nuclease digests circular RNAs at a much slower rate. Therefore, the treatment has to be administered carefully. If the treatment is not long enough, many of the linear RNAs are not completely digested. If the treatment is too long, the circRNAs are degraded as well. The following section will highlight the results obtained from *RNaseR* treated libraries.

RNaseR treatment enriches for circular RNA

Mapping these long reads in the described two-step approach (1. all reads as paired-end; 2. all unmapped reads as single-end) resulted in an 85 % mapping rate in both heart samples and a 60 % mapping rate in both liver samples. The proportion of chimerically mapped reads ranged from 10 % in the young liver to 20 % in the old liver. However, these numbers were not reflected in the circRNA diversity and abundance, where both liver samples had similar amounts of detected circRNAs on all confidence levels (support of at least 2, 4, or 10 reads). Overall, the heart had fewer circRNAs detected but the number of high confident circRNAs ($\geq$ 10 supporting reads) exceeded the number of high confident circRNAs in liver by two to three times. This was also reflected by the number of reads per back-splice junction in Figure 2.7c. The median number of reads per junction in heart was three times as high as in liver, thus the heart had a less diverse circRNA landscape. However, these circRNAs were more abundant.

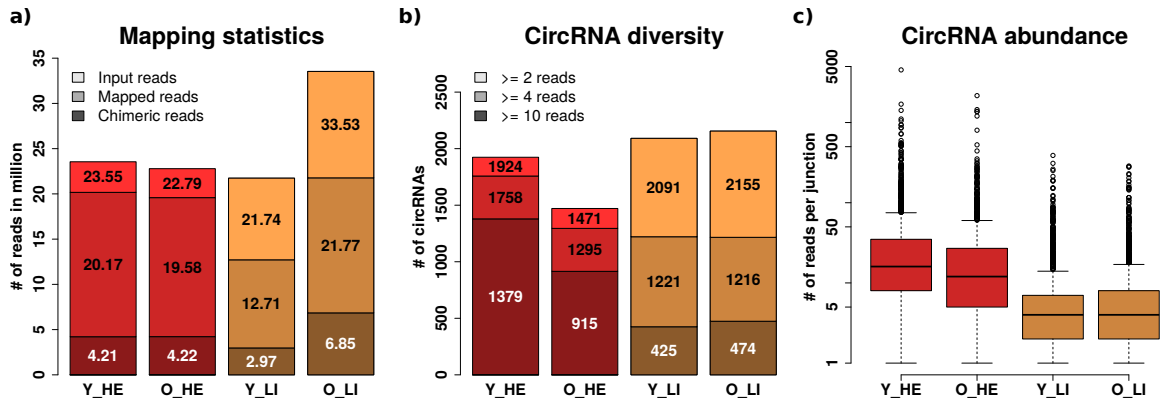

Figure 2.7: Summary statistics a) Mapping statistics; the barplot is additive but the numbers noted on in the bars are summarized appropriately. **b)** Circle diversity; the barplot is additive and the numbers in the bars represent the summarized number of junctions with at least x reads supporting the circular junction. **c)** Circle abundance; boxplot summarizes the number of reads supporting each identified circular RNA

Core circRNA are indicated in regulation of metabolic processes

In contrast to Figure 2.7, where the circRNA landscape is viewed for each sample individually, Figure 2.8 compares the circRNA landscape between samples. This comparison shows different circRNAs expressed in the young and old heart while the same circRNAs are expressed in the young and old liver. The heart had a core set of 1218 circRNAs. Comparing the host genes of these circRNAs

to all expressed genes in a GO enrichment analysis revealed that their host genes were enriched for lipid modification ($q = 0.006$). The liver had a core set of 1986 circRNAs. A GO enrichment of these core circRNAs indicated that the host genes of these circRNAs were enriched for cellular catabolic processes ($q = 7.60e-06$). Both core sets were enriched for various metabolic processes, such as regulation of macromolecule metabolic processes ($q = 1.82e-13$) and regulation of primary metabolic processes ($q = 4.80e-12$, see Table 2.4 for details).

The four way Venn diagram, comparing all samples at once, shows the young samples sharing 1366 circRNAs and the old samples sharing 1042 circRNAs. This suggests that the sample O_HE is the limiting factor in the analysis. However, it remains elusive whether the circRNA diversity truly decreases with age in the heart and remains stable in the liver. It could also be an effect that was only seen in the analysed samples, as one sample for each time point and tissue is not appropriate for any statistics.

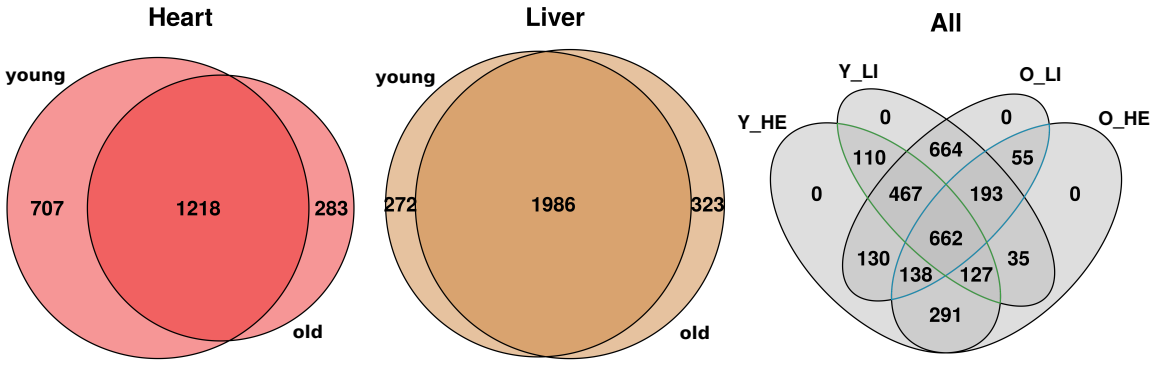

Figure 2.8: Agreement of samples. Heart samples only share 1218 circRNAs. Ninety percent of the circRNAs are shared between young and old liver. Comparing all samples reveals a set of 662 core circRNAs shared among all circRNAs. The sum of the green encirclement corresponds to the amount of circRNAs shared between the young samples (1366). The sum of the blue encirclement corresponds to the amount of circRNAs shared between old samples (1042).

Table 2.4: GO enrichment of genes expressing circRNAs

Enriched in	GO Term	FDR	Enrichement
Both	regulation of macromolecule metabolic process	1.82e-13	1.36
	regulation of metabolic process	2.16e-13	1.34
	regulation of cellular metabolic process	4.41e-12	1.33
	regulation of primary metabolic process	4.80e-12	1.34
	biological regulation	4.90e-12	1.19
	regulation of biological process	1.49e-11	1.20
	cellular macromolecule metabolic process	9.20e-11	1.30
	regulation of cellular process	8.18e-11	1.20
	regulation of nitrogen compound metabolic process	1.06e-10	1.39
	cellular process	2.68e-10	1.13
Heart	lipid modification	0.00638	2.47
	negative regulation of epithelial cell proliferation	0.00684	2.71
	glial cell differentiation	0.01900	3.23
	response to radiation	0.02180	1.84
	regulation of neurotransmitter secretion	0.02430	3.53
Liver	cellular catabolic process	7.60e-06	1.57
	positive regulation of RNA biosynthetic process	6.84e-05	1.50
	positive regulation of nitrogen compound metabolic process	8.60e-05	1.42
	positive regulation of transcription, DNA-templated	1.26e-04	1.48
	positive regulation of nucleic acid-templated transcription	1.24e-04	1.48

2.3.4 Conclusions and outlook

With the combination of *RNaseR* treatment and long reads (2 x 250 BP), this new dataset provides a powerful resource to study the circular RNA landscape. An analysis of the estimated circRNA length, including only annotated exons, revealed that heart circRNAs were 100 BP longer than liver circRNA (520 BP vs 433 BP).

Evaluating the sequence length of read pairs that span the back-splice junction revealed that the length of chimeric reads spanned a wide range from 17 BP to 249 BPs. This could be explained by the way chimeric reads are saved in an alignment file. They are split at the back-splice junction and are saved as two fragments. Only the mates that do not span the back-splice junction remain intact, i.e have a fragment length of 249 BP.

Evaluating the insert size (see Figure 2.9 for details) revealed that the insert length in heart ranged from shorter 200 BP inserts to long 600 BP inserts, while inserts in liver were much shorter, ranging from 300 BP to 400 BP. Again, the large amount of inserts shorter than 50 BP could be explained by the fragmentation of the read during the chimerical read mapping.

Correlating the circle length and insert size (Figure 2.10), it became evident that this dataset could provide a powerful resource to fully capture short circles. Thus, the exon-intron structure of short circular RNAs could be fully reconstructed, and the exon-intron structure of medium length circular RNAs could be partially reconstructed. This new knowledge would have many implications for further analyses such as the validation of circRNAs, the overexpression, or knockdown of circRNAs. More importantly, it could provide the possibility to perform more accurate computational analysis, such as miRNA seed and RBP motif enrichment.

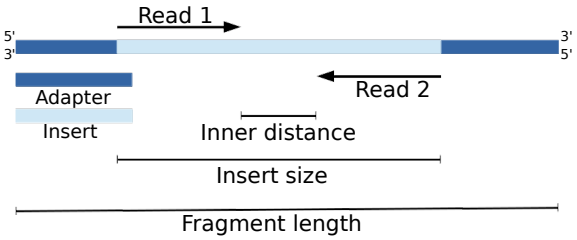

Figure 2.9: Schematic view of the insert size. Paired-end RNA-Seq library characteristics are described by their inner distance between the two reads, insert size (inner distance + sequenced reads without adapter), and fragment length (length of the sequenced fragments including adapters)

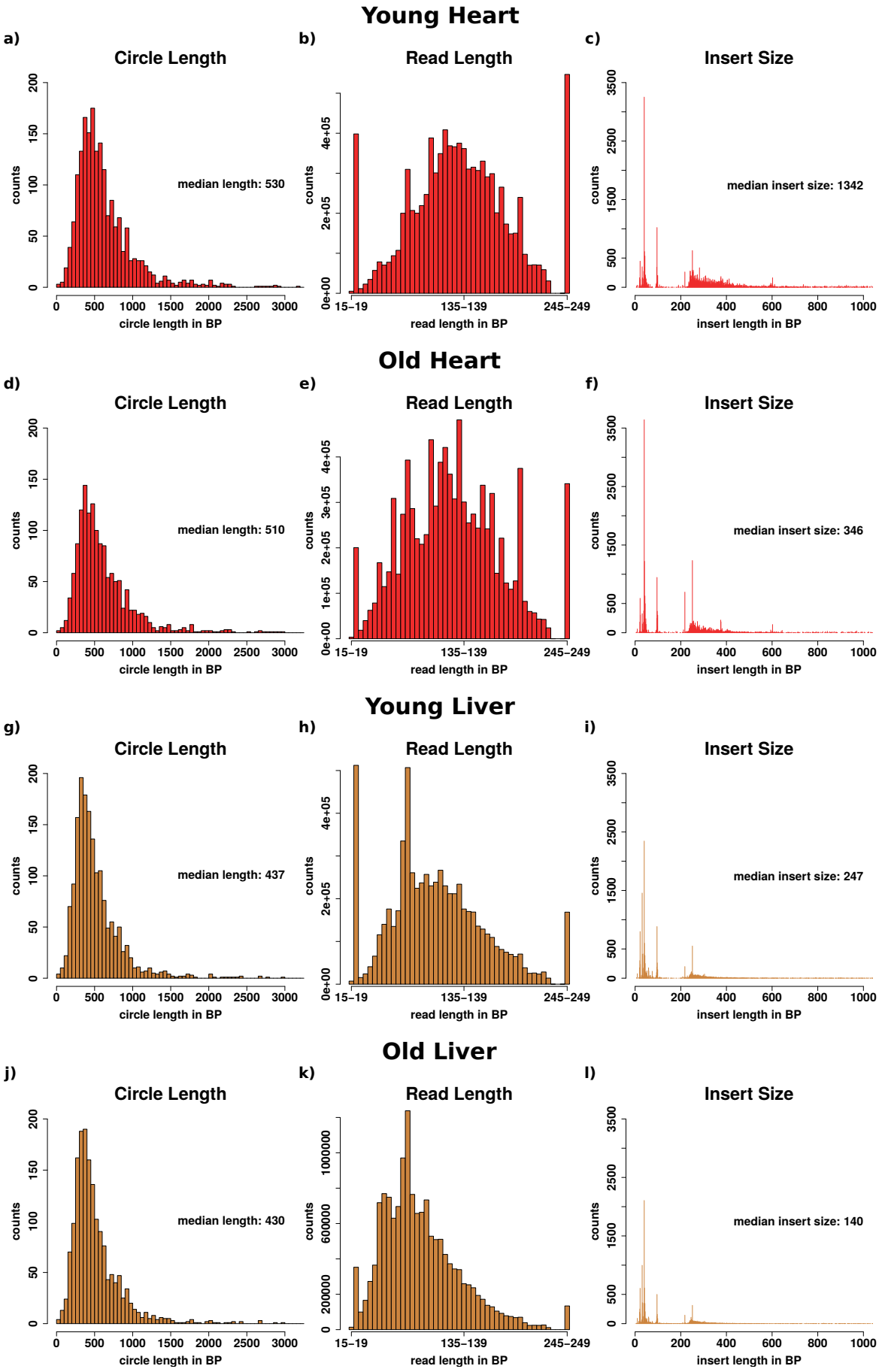

Figure 2.10: Overview of the circRNA and library characteristics.

Chapter 3

Towards full circular RNA characterization using sequencing data – FUCHS

3.1 Introduction

Alternative splicing (AS) is a well known and documented phenomenon in linear mRNAs, but understudied in circRNAs. The major problem studying AS in circular RNA is to reliably discern alternative splicing isoforms from circRNAs to those from mRNAs. Studying alternative splicing in circular RNAs could lead to crucial implications for validation and deciphering the function of circRNAs.

Implications for Validation

Circular junctions can be validated by performing a polymerase chain reaction (PCR) using outward facing primers. If the gene of interest forms only one circRNA, the PCR will have one specific band. If on the other hand the gene of interest forms several different circRNAs, a smear or several bands will be formed (see Figure 3.1 for details). Not knowing how many circRNAs are expressed by a

gene can greatly influence the validation process. Some candidates could be miss classified as false positives due to smears or seemingly unspecific bands.

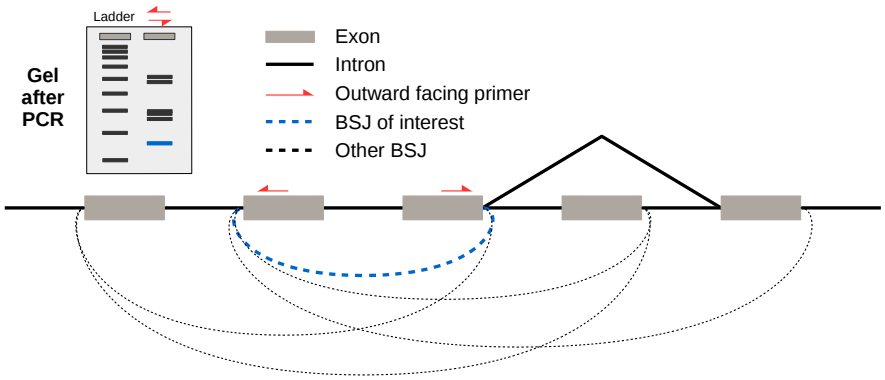

Figure 3.1: Problems for circRNA validation by running a gel electrophoresis. A shematic view of a host gene expressing several different circRNAs. These circRNAs could be low abundant circRNAs and thus not detected by RNA sequencing. Trying to validate the only detected back-splice junction (portrayed in blue) with PCR and gel electrophoresis would lead to several bands on the gel, thus might be falsely dismissed as false positive.

Deciphering the function of circRNAs

The knowledge about genes that express more than one circRNA, especially internal AS isoforms, can be used in computational functional analysis of circRNAs. A motif search in flanking introns of frequently circularized exons compared to less frequently circularized exons might lead to unknown RNA binding proteins or transcription factors (TF) responsible for the circle formation. Performing a miRNA seed search on frequently circularized exons over less frequently circularized exons might provide an indication if there is a purpose for the frequent circularization of one exon. If there is one exon which is highly enriched for miRNA seeds, maybe it does not matter which other exons are included in the circRNA.

3.1.1 State of the art programs

The circRNA detection tools, mentioned in chapter 2, are only able to discern circRNAs if they do not share the same back-splice junction. Two of these programs recently published upgrades in which they attempt to identify alternatively spliced isoforms in circRNAs.

CIRCexplorer2

CIRCexplorer2, by the Yang lab at the Shanghai Institutes for Biological Sciences, adapts several measurements implemented for alternative splicing in linear RNAs. These are PSU, percent splice-site usage, to identify alternative 3' and 5' splice sites; PSI, percent spliced in, to identify skipped exons; and PIR, percent intron retention, to identify retained introns (see Figure 3.2a). All these measures are calculated based on $polyA^-$ and/or $polyA^-/RNaseR^+$ libraries, assuming that all linear RNAs were degraded using *RNAseR* treatment. Comparing circular isoforms to their linear cognates (AS detected from $polyA^+$ libraries) they identify the predominant form (circular or linear) of each isoform.
A possible drawback is that they make assumptions about the efficiency of the library preparation protocols which is accepted not to be 100 % efficient, meaning that there will be circular RNA in the $polyA^+$ fraction and linear RNAs in the $polyA^-/RNaseR^+$ libraries. Furthermore, to detect circRNAs using this tool requires RNA-Seq of at least $polyA^-$ as well as $polyA^+$ sequencing. As we only generated $RNaseR^+$ libraries, this program is not suited to study AS in circRNAs in the aforementioned mouse dataset.

CIRI-AS

The developers of CIRI, one of the programs benchmarked in the chapter 2, recently developed a program to identify alternative splicing within circRNAs. Contrary to CIRCexplorer2, they require only one sample to detect alternative splicing and support $RNaseR^+$ as well as $RNaseR^-$ libraries. First, circRNAs are identified by finding back-splice junction reads using CIRI. Second, forward-splice junctions are analyzed within these back-spliced read pairs using CIRI-AS. For circRNAs without forward-spliced reads, CIRI-AS estimates the internal structure based on linear reads. The forward-spliced junctions are then sorted and all possible routes from one anchor of the back-splice junction to the other anchor are noted (see Figure 3.2b). After filtering, the remaining routes between the same anchors are classified into four categories: alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), exon skipping (ES), and intron retention (IR). They found ES to be the most prominent alternative splice form, while intron retention accounted for only 1% of alternatively spliced events (see Figure 3.2c).

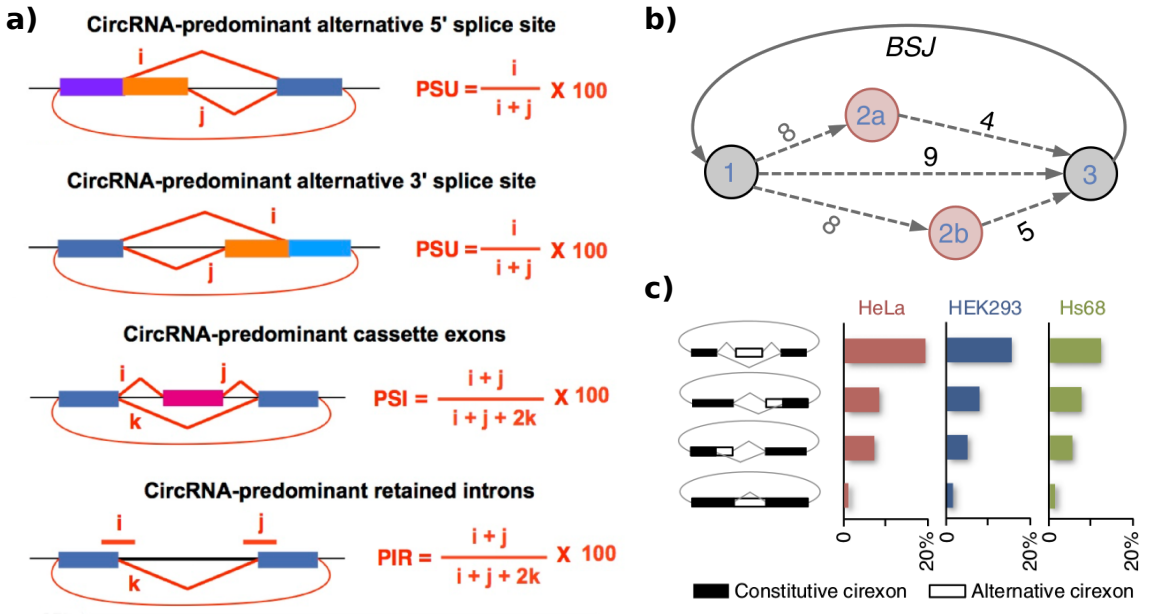

Figure 3.2: State of the art programs on alternative splicing in circRNAs. **a)** CIRCexplorer2 classifies alternative splicing into four categories and calculates specific scores for each (modified from [45]). **b)** Shematic view of the approach taken by CIRI-AS to identify alternative splicing in circRNAs. **c)** CIRI-AS also classifies alternative splicing into four categories and shows the distribution of each type of alternative splicing among highly abundant circRNAs (> 20 BSJ reads). **b** and **c** are modified from [46].

FUCHS

When I started working on circRNA, most computational prediction pipelines used RNA-Seq reads only to identify back-splicing events. My pipeline FUCHS: **FU**ll **CH**aracterization of circular RNA using RNA-**S**equencing; extends this concept by considering all RNA-Seq information from long reads (typically $>$ 150 bp), to further expand our understanding of possible roles of circular RNAs. I took a similar approach as in CIRI-AS, but using STAR/DCC results as I previously showed that it outperforms BWA/CIRI.

Besides alternative splicing in circRNAs, by running FUCHS, the user will also learn more about the exon coverage, the amount of double-breakpoint fragments, and the different circular isoforms arising from one host-gene. This new knowledge will enable the user to perform differential motif enrichment and miRNA seed analysis to determine potential regulators during circRNA biogenesis. FUCHS is an easy-to-use Python based pipeline that contributes with new aspects to the field of circRNA research.

3.2 The pipeline

A flow diagram of the main steps executed by FUCHS is visualized in Figure 3.3. FUCHS is a tool designed as a Python pipeline to address several questions after circRNA candidates were identified. The following section will discuss the different features of FUCHS.

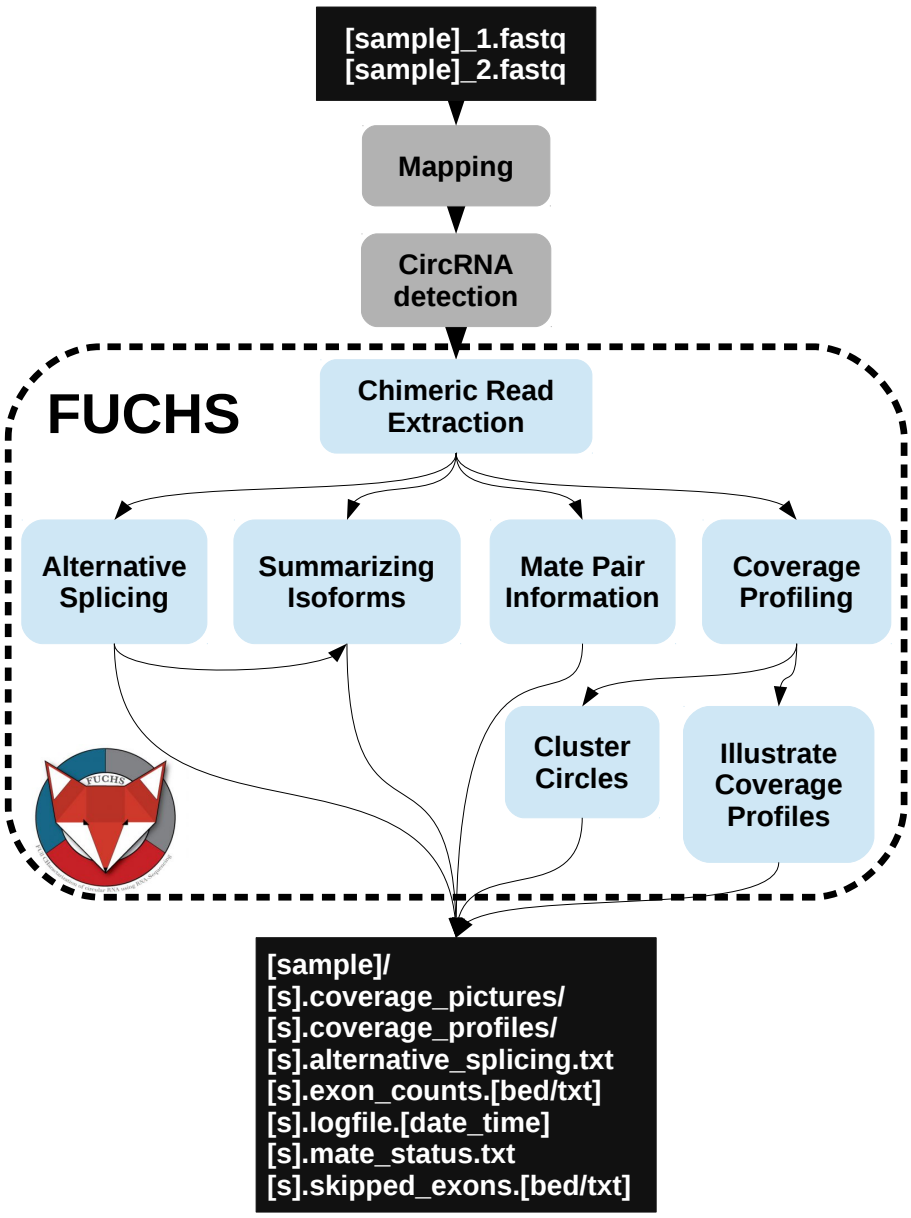

Figure 3.3: Schematic view of the FUCHS workflow. The black boxes represent input and output files. The gray boxes represent steps that any user has to perform before running FUCHS. The light blue boxes are steps which are started by running FUCHS. The arrows represent dependencies between the steps. The logo represents a fox on a circRNA (credits Logo: Anne Ammerstorfer)

3.2.1 Input data

To run FUCHS, the user has to first map sequencing reads and detect circRNAs, for example using STAR to map the reads and DCC to detect the circRNAs as described in chapter 2.3.

The pipeline requires only three input files. First, a tab-separated list of circles, where the first column contains circleIDs and the second column contains a comma separated list of names of reads spanning the chimeric junction (see Table 3.1). Second, a BAM or SAM file containing all chimerically mapped reads, which may also contain linearly mapped and unmapped reads (see Table 3.2). Third, a BED formatted annotation file (see Table 3.3) to identify skipped exons and describe the exon-usage. If circRNAs were detected by DCC, the first file is not needed. FUCHS is able to extract all necessary information from the Chimeric.junction.out files produced by STAR and CircRNACount produced by DCC.

Table 3.1: Input circID file

CircID	Read names
5:92030983|92033757	SN863:53528,SN863:23096,SN863:76567,...
6:39357400|39367766	SN863:76194,SN863:62105,SN863:96141,...
7:41611676|41613251	SN863:93309,SN863:69105,SN863:97405,...
...	...

Table 3.2: Input BAM file

Read name	FLAG	Chr	Start	MAPQ	CIGAR	...	breakpoints
...							
SN863:76567	16	5	92030983	3	54S146M1995N49M	...	jI:B:i,92031129,92033123
SN863:76567	272	5	92033715	3	11S43M195S	...	jI:B:i,-1
SN863:76567	0	5	92030983	3	44S146M1995N48M11S	...	jI:B:i,92031129,92033123
SN863:76567	256	5	92033715	3	1S43M205S	...	jI:B:i,-1
...							
SN863:96141	16	6	39357400	3	76S84M2466N89M	...	jI:B:i,39357484,39359949
SN863:96141	272	6	39367691	3	76M173S	...	jI:B:i,-1
SN863:96141	0	6	39367000	3	47M597N123M79S	...	jI:B:i,39367047,39367643
SN863:96141	256	6	39357400	3	170S79M	...	jI:B:i,-1
...							

Table 3.3: Input reference gene model

Chr	Start	End	Name	Score	Strand
...					
5	92024935	92025024	ENSMUST00000069937_cds_6	0	-
5	92030982	92031128	ENSMUST00000069937_cds_7	0	-
5	92033123	92033236	ENSMUST00000069937_cds_8	0	-
5	92033578	92033757	ENSMUST00000069937_cds_9	0	-
5	92037167	92037362	ENSMUST00000069937_cds_10	0	-
...					
6	39354976	39355122	ENSMUST00000090243_cds_9	0	-
6	39357399	39357483	ENSMUST00000090243_cds_10	0	-
6	39359949	39360042	ENSMUST00000090243_cds_11	0	-
6	39364502	39364611	ENSMUST00000090243_cds_12	0	-
6	39366888	39366977	ENSMUST00000090243_cds_13	0	-
...					

If circRNAs were detected using STAR and DCC, the user should merge reads from the chimeric alignment files using the following commands.

```
samtools view -Sb -o mate.1.bam mate1.Chimeric.out.sam

samtools view -Sb -o mate.2.bam mate2.Chimeric.out.sam

samtools view -Sb -o full.bam Chimeric.out.sam


samtools sort -o mate.1.sorted.bam mate.1.bam

samtools sort -o mate.2.sorted.bam mate.2.bam

samtools sort -o full.sorted.bam full.bam


samtools merge sample.bam mate.1.sorted.bam mate.2.sorted.bam full.sorted.bam
```

3.2.2 Running the pipeline

Once the input is prepared, all steps of FUCHS are started with one of the commands outlined below.

```
# with STAR/DCC
FUCHS -r [NUM] -q [NUM] -s [CHAR] -p [ensembl|refseq] -e [NUM]
      -T [tmp/folder] -D CircRNACount -J Chimeric.out.junction
      -F mate1.Chimeric.out.junction -R mate2.Chimeric.out.junction.fixed
      -B sample.bam -A [annotation].bed -O [FUCHS/output] -N [sample_name]
```

```
# if circID file is present
FUCHS -r [NUM] -q [NUM] -s [CHAR] -p [ensembl|refseq] -e [NUM]
      -T [tmp/folder] -C circID -B sample.bam -A [annotation].bed
      -O [FUCHS/output] -N [sample_name]
```

Where `-r` sets the read support threshold for circRNAs to be considered in the analysis. `-q` sets the quality threshold of these reads, e.g. if a circRNA has ten reads with `MAPQ` = 1 and only one read with `MAPQ` = 3, the circRNA would not be considered; while a circRNA with two reads and `MAPQ` = 3 will be considered for further analysis. The number of reads is counted as number of unique read names. `-s`, `-p`, and `-e` are important to correctly parse the gene name and exon index from the annotation file. `-s` indicates the character by which the name string is separated. `-p` indicates if the gene name comprises of only the first field as in ENSEMBL annotation or the first two fields as in RefSeq annotation. `-e` specifies the field listing the exon index. For example, to parse `ENSMUST00000132064_cds_10_0_chr1_8624779_r` parameters need to be set as the following: `-p ensembl`, `-s _`, and `-e 2`; to parse `NM_012102_exon_10_0_chr1_8365812_r` parameters need to be set as the following: `-p refseq`, `-s _`, and `-e 3`.

All steps of the pipeline can also be run individually by importing and creating objects of the classes as I will describe in the following sections.

3.2.3 Extracting chimeric reads: `extract_reads.py`

It is often a good practice to verify the alignments of predicted circRNAs as a first quality assessment of the data before proceeding with the downstream analysis. The identification of chimerically mapped reads by eye is nearly impossible because these reads are mixed with linearly mapped reads or other circRNAs originating from the same locus. Therefore, `FUCHS` starts by separating chimerically mapped reads into individual BAM files such that one BAM file contains only chimeric reads which span the same back-splice junction (see Figure 3.4). The script `extract_reads.py` requires a list of circRNAs with affiliated reads and a BAM file containing at least all chimeric reads as well as a folder path and sample name for the circle BAM files to be written to. The list of all chimeric reads is loaded into Python, the BAM file is loaded using `pysam`, and all chimeric reads, which are in the list of chimeric

reads, are saved. For each circRNA, which is passing a user defined threshold of at least `r` reads with a mapping quality of at least `q`, the chimeric reads are written into a BAM file. The file name is defined as follows: `[chr]_[circle start]_[circle end]_[# of reads]reads.bam`. All BAM files are then sorted and indexed using `pysam`. Besides the improved visibility of chimeric reads in the genome browser, parallelizing downstream steps becomes straightforward, thus enhancing the performance of `FUCHS`.

The user can also run extract chimeric reads within Python by creating an object of class `extract_reads`:

```
import extract_reads as extract_reads
er = extract_reads.extract_reads(cutoff_reads, cutoff_mapq,
     circles, bamfile, outfolder, sample, tmp_folder)
er.run()
```

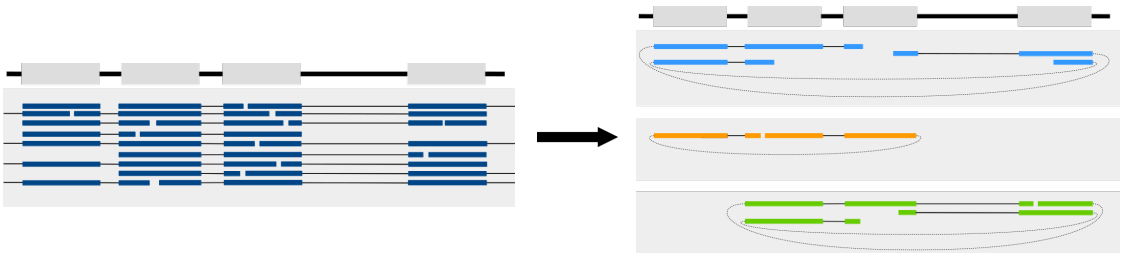

Figure 3.4: Chimeric read extraction in `FUCHS`. On the left-hand side of the figure is the alignment file of all reads. It is impossible to see which reads belong to which circRNA. The `extract_reads` method extracts chimeric reads and seperates them into individual BAM files as pictured on the right-hand side of the figure.

3.2.4 Alternative Splicing: `detect_skipped_exons.py`

Alternative splicing is a well known mechanism to increase the complexity of the transcriptome. Alternative splicing could also increase the diversity in the circular RNA landscape. Knowing all AS events within one circRNA can be crucial when validating circRNA by qPCR or deciphering potential functions of circRNA. The script `detect_skipped_exons.py` identifies skipped exons in `FUCHS`. It requires a circle BAM file and a gene model BED file as input. First, all linear introns are identified from the chimerically split reads. Subsequently, the intron coordinates are intersected with annotated exons using `pybedtools.intersect()`. Any exon that overlaps with an intron is

considered as a skipped exon candidate. To evaluate if the candidate exon is an alternatively skipped exon, it is intersected with the circle BAM file and the number of reads aligning to this exon are counted. If there is at least one read aligning to the candidate skipped exon, it is reported in the [sample].skipped_exons.[txt/bed] file with the ratio of reads aligning to the skipped exons vs. reads skipping the exon. The text file also lists all reads skipping the candidate exon to better trace the signal.

Table 3.4 shows an example output of the script `detect_skipped_exons.py` which may be executed individually within Python by creating an object of class `detect_skipped_exons`:

```
import detect_skipped_exons as skipped_exons
se = skipped_exons.detect_skipped_exons(outfolder, sample,
     bedfile, tmp_folder, platform)
se.run()
```

Table 3.4: sample.skipped_exons.txt

circle_id	transcript_id	skipped_exon	intron	read_names	splice reads	exon reads
2_92214049_92230724	ENSMUST162146	2:92230637-92230724	(2,92228482,92230658)	SN863:80759,..	7	32
7_89955531_89962361	ENSMUST107234	7:89956254-89956413	(7,89955605,89956343)	SN863:5259,..	2	4

3.2.5 Isoform Summary: `detect_splicing_variants.py`

To better evaluate the differences in circRNA diversity between samples, one needs to know how many circRNAs are expressed by one host gene and in what relationship they are to each other. To summarize circRNAs based on their host gene, the script `detect_splicing_variants.py` requires the following input files: a list of circRNA coordinates and an annotation file. The program generates a host gene based table classifying circRNAs from the same host gene into four different categories: same_start, same_end, overlapping and within (see Figure 3.5 for a graphical representation of circRNA relationships).

To summarize only the isoforms within Python the user may create an object of class `detect_splicing_variants`:

```
import detect_splicing_variants as splicing_variants
sv = splicing_variants.detect_splicing_variants(split_character,
     platform, circles, bedfile, outfolder, sample, tmp_folder)
sv.run()
```

Table 3.5 illustrates and example output table of `detect_splicing_variants.py`.

Table 3.5: sample.alternative_splicing.txt

Transcript	circles	same_start	same_end	overlapping	within
ENSMUST22222	circ_1,circ_2	circ_1|circ_2,	.	.	.
ENSMUST98519	circ_1,circ_2	.	.	.	circ_1|circ_2,
ENSMUST42561	circ_1	.	.	.	.
ENSMUST93769	circ_1,circ_2,circ_3	circ_1|circ_2,	circ_2|circ_3,	.	.

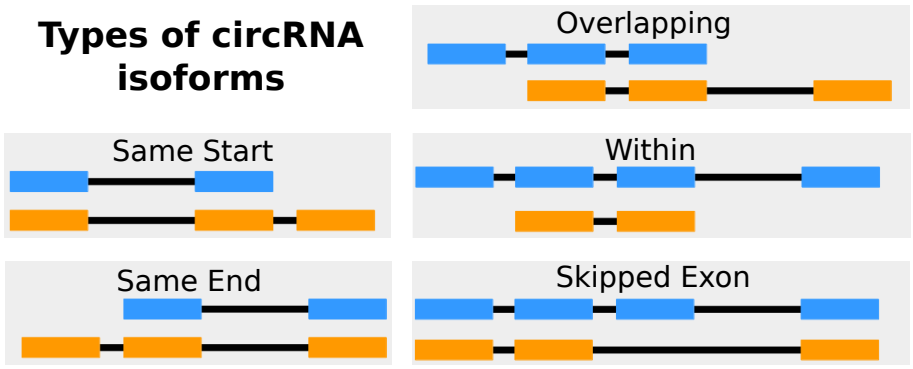

Figure 3.5: Isoforms of circRNAs from one host gene. Graphical representation of possible relationships between two circRNAs. Same start, same end, overlapping and within are counted by the method `detect_splicing_vairants`. Skipped exons are detected by the method `detect_skipped_exons`.

3.2.6 Mate-pair Information: `get_mate_information.py`

Back-spliced junctions may originate from other events besides circulatization. Events such as trans-splicing and tandem duplication also result in back-splice junctions. Assuming that circular RNAs are still intact during the cDNA synthesis the reverse transcriptase could continue to transcribe around the whole circRNA possibly synthesizing the back-splice junction twice (see Figure 3.6). These rolling circles have been observed by Matsumoto *et al.* and You *et al.* [47, 48]. Thus, double-breakpoint fragments could indirectly validate the circularization of the sequenced exons. Based on the circle BAM files, the script `get_mate_information.py` counts how often only one mate of a read pair spans the back-splice junction, and how often both mates of the read pair span the back-splice junction.

The output is a tab-separated table indicating the number of single and double-breakpoint fragments as well as the minimum circle length (shortest possible transcript) and maximum circle length (longest possible transcript, or circRNA length from start to end, if no annotation overlaps the circRNA coordinates) because rolling circles are only possible if the circle length is shorter than the synthesized fragment length.

The user may execute this step within python by generating an object of class `get_mate_information`:

```
import get_mate_information as mateinformation
mi = mateinformation.mate_information(platform, split_character,
     bedfile, outfolder, sample, tmp_folder)
mi.run()
```

Table 3.6 shows an example output produced when running `get_mate_information`.

Table 3.6: sample.mate_status.txt

circle_id	transcript_ids	num_ reads	min_ length	max_ length	single	double	undefined
10_108240105_108251008	ENSMUST070663	4	757	757	1	3	0
10_117692582_117695068	ENSMUST105263,ENSMUST020408	12	90	319	1	11	0
10_120947204_120952255	ENSMUST119093,ENSMUST119944	22	172	172	0	22	0

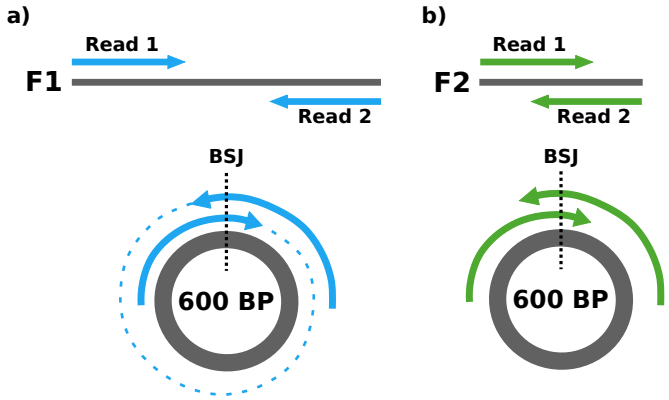

Figure 3.6: Double-breakpoint fragments. Images of two scenarios of how double-breakpoint fragments may originate from circular RNA. The fragment on the left-hand side results from a rolling circle amplification spanning the back-splice junction twice. Reads originating from this fragment both span the chimeric junction, but on different locations on the fragment. The fragment on the right-hand side is shorter than on the left. Reads originating from this fragment overlap in the middle, thus both reads span the chimeric junction, but it is actually the same location on the fragment. In RNA-Seq data it is impossible to distinguish the two cases.

3.2.7 Coverage Profiling: `get_coverage_profile.py`, `make_coverage_picture.R`, and `summarized_coverage_profiles.R`

The coverage profile of a circRNA is important for many reasons. Because all reads considered as circular RNAs are anchored at the back-splice junction, a typical profile of a circRNA will be an inverted bell shape with high coverage around the circle junction boundaries and a low or no coverage in the middle, i.e. opposite the back-splice junction. Thus, the coverage profile can indicate how much of a circRNA can be reconstructed (no gap in the coverage profile indicates that the structure of the circRNA can be fully reconstructed). An uneven coverage profile might indicate a bias in sequencing, alignment, or annotation and this circRNA should be investigated further to ensure that it is a true circRNA. The script `get_coverage_profile.py` uses the coverage function from `pybedtools` to generate a position-wise as well as an exon-wise coverage profile based on the circle BAM files and the annotation BED file.

Table 3.7 is an example of the position-wise coverage profile. Each profile is written to a separate file while the exon-wise coverage profiles are written into one BED formatted table to get an overview of the exon coverage (see Table 3.8). The user may generate coverage profiles directly from separated circle BAM files by creating an object of class `get_coverage_profile`:

```
import get_coverage_profile as coverage_profile
sv = coverage_profile.get_coverage_profile(exon_index, split_character,
     platform, bedfile, outfolder, sample, tmp_folder)
sv.run()
```

Table 3.7: Position-wise coverage track

exon	relative_pos_in_circle	relative_pos_in_exon	coverage
8	1	1	14
8	2	2	14
8	3	3	14
8	4	4	14
...			

Based on the position-wise coverage profiles the R script `make_coverage_picture.R` generates a graphical interpretation of the coverage. The profiles are smoothed and exons are indicated by different colors. The graphs are saved as PNGs such that the user may easily view them side by side

Table 3.8: Exon-wise coverage track

sample	circle_id start fragments	transcript_id end number+	other_ids strand number-	exon_id exon_length	chr unique_reads
old_heart	1:160700712-160702460 160700711 100	ENSMUST28049 160700878 50	ENSMUST28049 - 50	11 167	1 50
old_heart	1:160700712-160702460 160702357 100	ENSMUST28049 160702460 50	ENSMUST28049 - 50	12 103	1 50
old_heart	9:59394878-59405855 59394877 2	ENSMUST171975 59394990 0	ENSMUST171975,ENSMUST165322 - 2	1 113	9 2

or scan through them to evaluate the quality and coverage of the circRNAs. The user may generate coverage pictures by executing the R script:

```
make_coverage_picture.R [sample].coverage_profiles/circle.BAM
[sample].coverage_pictures/
```

A second R script (`summarized_coverage_profiles.R`) summarizes all circRNA coverage profiles. After normalizing by length, the circRNAs are clustered based on their coverage profile similarities. The clustering is implemented on different sets of circRNAs; once on all circRNAs to obtain a general overview and once on circRNAs shorter than 500 BP, on circRNAs between 500 to 1000 BP, and circRNAs longer than 1000 BP to avoid biasing the clustering to only cluster circRNAs by their length. K-means from the R package `amap` is implemented as method for the clustering. The number of cluster starts is chosen based on the following rule: if a set contains two or fewer circRNAs, no clustering is performed. If a set contains three to nine circRNAs, k = 2. If a set contains between ten and 20 circRNAs, k = 3. If a set contains 20 to 100 circRNAs, k = 4. If a set contains more than 100 circRNAs, the number of clusters is determined by k = *round* $\left(\frac{\sharp \text{ circRNAs}}{20}\right)$. Independent of the size of the sets, the maximum number of clusters chosen is ten. Running the script requires a set of input coverage profiles, which are read into R from a user specified input folder, and results in tab separated tables (see Table 3.9) indicating the cluster a circRNA belongs to as well as tables (see Table 3.10) indicating the cluster means. The user may only cluster the circRNAs based on the coverage profiles by executing the R script:

```
summarized_coverage_profiles.R [sample].coverage_profiles/
```

Table 3.9: cluster_association.all_circles.tsv

circle_id	length	cluster_id
10_108240105_108251008	442	8
10_12436281_12455564	202	1
10_128207298_128208667	229	6

Table 3.10: cluster_means.all_circles.tsv

	X1	X2	X3
1	5.61	5.40	4.80
2	8.25	8.25	8.5
3	1.86	1.90	1.96
⋮			

3.2.8 Output data

Running the whole pipeline results in three folders and seven files in the specified output folder. Table 3.11 provides an overview over the whole output obtained by running FUCHS in the described way.

Table 3.11: Output files

File	Script	Reference
[sample]	extract_reads.py	
[s].coverage_pictures	make_coverage_picture.R	
[s].coverage_profiles	get_coverage_profile.py	
[s].alternative_splicing.txt	detect_splicing_variants.py	Table 3.5
[s].exon_counts.[bed|txt]	get_coverage_profile.py	Table 3.8
[s].logfile.[date_time]	FUCHS	
[s].mate_status.txt	get_mate_information.py	Table 3.6
[s].skipped_exons.[bed|txt]	detect_skipped_exons.py	Table 3.4

3.3 Results and discussion

Alternative splicing and isoform summary

FUCHS ran on a subset of circRNAs detected by DCC based on the read threshold of at least two reads having a mapping quality of at least 2 (`-r 2 -q 2`). Figure 3.7a illustrates the number of circRNAs per sample, categorized by length, that were analysed. It appears that though, DCC detected more circRNAs in both liver samples than in both heart samples (compare Figure 2.7), there were more circRNAs that did not pass the threshold in liver than in heart ($\sim$ 300 in liver vs. $\sim$ 30 in heart). Summarizing these circRNAs by host genes revealed that the majority ($\sim$ 60 %) of host genes expressed only one circRNA. CircRNAs that originated from the same host gene often shared either

the 5' splice site or the 3' splice site. The remaining three relationship categories were less frequent (3-7 %, see Figure 3.7c and Table 3.12). For host genes expressing more than two circRNAs each pairwise comparison was categorized. For example, if there were four circRNAs from the same host gene, three were using the same 5' splice site, and the fourth was sharing the 3' splice site with one of the other three this host gene would contribute one count to the 'same start' category and one count to the 'same end' category.

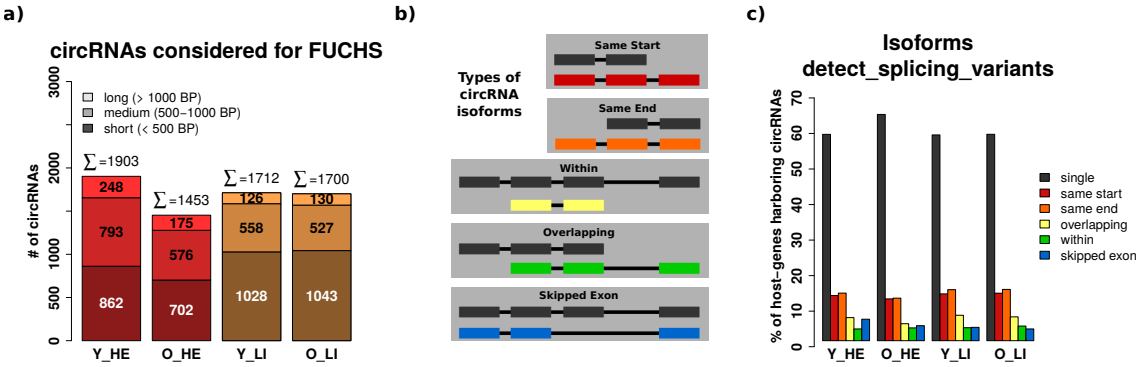

Figure 3.7: Isoform summary. a) Number of circRNAs analyzed with FUCHS in each sample stratified by the length of the circRNAs. The circle length was estimated based on the annotated gene model. **b)** Possible relationships between two circRNAs from one host gene. **c)** Proportion of relationshipts between two circRNAs from one host gene across all samples (see Table 3.12).

Table 3.12: Isoform summary

	Y_HE	O_HE	Y_LI	O_LI
single	1019	920	1125	1143
same start	223	170	256	263
same end	235	173	279	284
overlapping	114	69	139	132
within	58	52	71	81
skipped exon	106	61	73	65

Number of circRNAs per host gene correlates with host genes genomic length

More than 50 % of host genes harbored only one circRNA. Based on combinatoric possibilities I expected that the number of exons (i.e. number of splice sites) would correlate best with the number of circRNAs. However, calculating the Spearman rank-correlation coefficient revealed that the number of circRNAs correlated best with the genomic length and only second best with the number of exons ($r_S = 0.25$ vs. $r_S = 0.17$). The transcriptome length had little influence on the number of circRNAs

expressed by one host gene ($r_S = 0.13$). This suggests that the intron's lengths are of importance for circularization. If the intron's and exon's characteristics were equally contributing to the number of circRNAs one host gene can express, the correlation between the number of circRNAs and the genomic length as well as the transcriptomic length should be similarly high. Indeed when only taking the length of introns (genomic length - transcriptomic length) into consideration, the correlation increased to $r_S = 0.27$. This indicates that longer introns are beneficial for producing more circRNAs. Previously circRNAs had only been viewed individually, where it was shown that long flanking introns drive circularization (over other introns) [34]. Here I indicate that of all host genes expressing circRNAs the ones with the longer introns give rise to more circRNAs that other host genes.

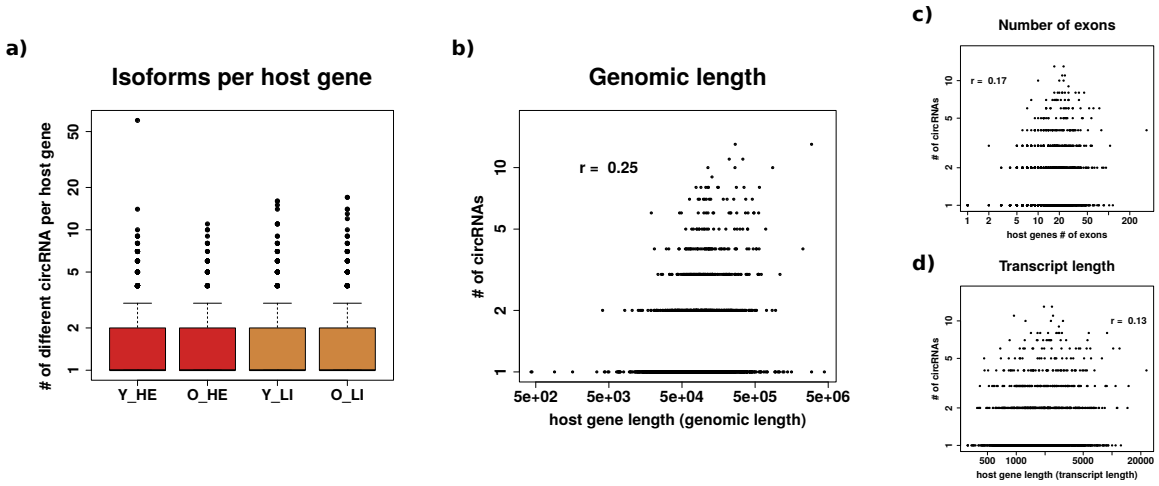

Figure 3.8: Correlation of number of circRNAs per host gene. a) Number of circRNAs per host gene for each sample. The majority of host genes expresses only one circRNA. **b)** Scatter plot comparing the genomic length of the host gene on the x-axis and the number of circRNAs expressed by the host gene on the y-axis. **c)** Number of exons annotated in the host gene on the x-axis and the number of circRNAs expressed by the host gene on the y-axis. **d)** Transcriptomic length (only exons) of the host gene on the x-axis and the number of circRNAs expressed by the host gene on the y-axis.

Mate pair information

The method `get_mate_information` provided information about the proportion of double-breakpoint fragments as well as an estimate about the minimal and maximal length of the circRNA. There was no significant difference in circle length when using the Wilcoxon test to compare young and old samples of the same tissue. However, when comparing heart circRNAs to liver circRNAs, heart circRNAs were significantly longer than liver circRNAs ($p < 2.2e-16$).

Based on the difference in circRNA length I expected that the proportion of double-breakpoint fragments would be higher in liver, assuming that short circRNAs could by fully captured during cDNA synthesis. Indeed, as visualized in Figure 3.9b, the proportion of double-breakpoint fragments was higher in liver than in heart.

A strong negative correlation ($r = -0.96$) between the circle length and the proportion of double-breakpoint fragments emerged when comparing the median circle length with the median proportion of double-breakpoint fragments as demonstrated in Figure 3.9c. However, it remains unclear if measuring the proportion of double-breakpoint fragments can really be used to identify false positive circRNAs as double-breakpoint fragments could also result from back-splice junctions spanning fragments shorter than the read length (see Figure 3.6). Nevertheless, some extremely short circRNAs showed single reads which actually span the back-splice junction twice. Unfortunately, these reads were not reported as such because current mapping algorithms are unable to map these reads correctly as they only allow for one chimeric junction per read.

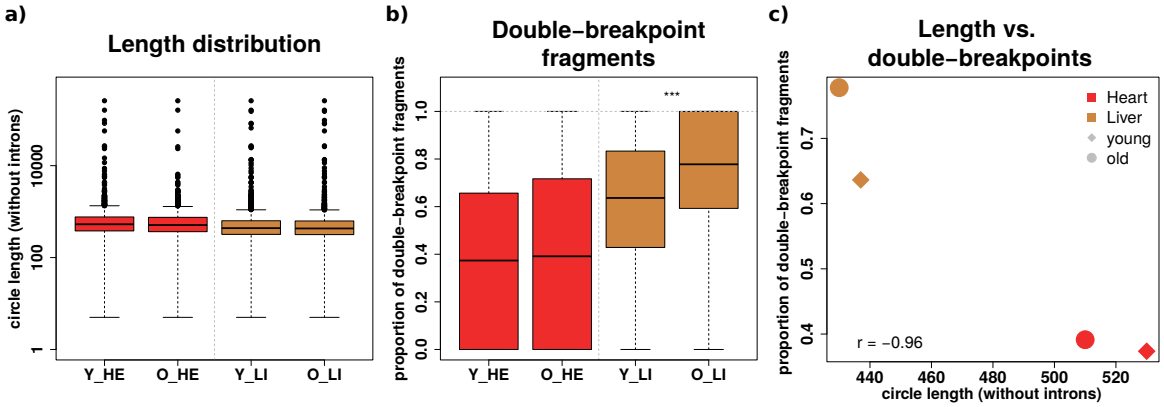

Figure 3.9: Double-breakpoint reads vs circle length. a) Distribution of estimated length of the circRNAs in each sample. **b)** Proportion of double-breakpoint fragments of the circRNAs in each sample. **c)** Median circle length for each sample on the x-axis and the median proportion of double-breakpoint fragments on the y-axis.

Coverage profiling

Figure 3.10 illustrates the average coverage profiles of all circRNAs grouped into three groups based on their length. As expected, short circRNAs were well covered with reads over the entire length. The longer a circRNA became, the wider the region with low or no support increases. This was true for all samples. Noticeably, though the gap grew wider in all samples, the junction coverage increased in the heart samples but not in the liver samples. This could mean that longer heart circRNAs were more abundant than shorter heart circRNAs.

Coverage profile clustering

Clustering all circRNAs based on their length-normalized coverage profile mainly separated all circRNAs by their different length (see Figure 3.11a and b). Cluster 3, 4, 7, and 9 indicated an unbalanced back-splice junction support. Cluster 4 and 7 only differ by the width of the gap, the same applied for cluster 3 and 9. This was explained by the length of their members. CircRNAs of cluster 3 and 4 were longer than those of cluster 7 and 9, while there was no difference between 3 and 4, 7 and 9 respectively.

By grouping circRNAs based on their length first and performing the clustering on the separated groups afterwards, a clustering solely by length was avoided as depicted with the box plots in Figure 3.11e, h, and k, thus suspicious circRNAs were more noticeable. I expect that the inner structure of circRNAs belonging to clusters 8 and 10 of short circRNAs, 8 and 9 of medium length circRNAs, and cluster 1 and 2 of long circRNAs will not match the annotated gene model rather than having a truly unbalanced junction support. This suggests that it may be worthwhile to describe the exon-intron chain in an unbiased way.

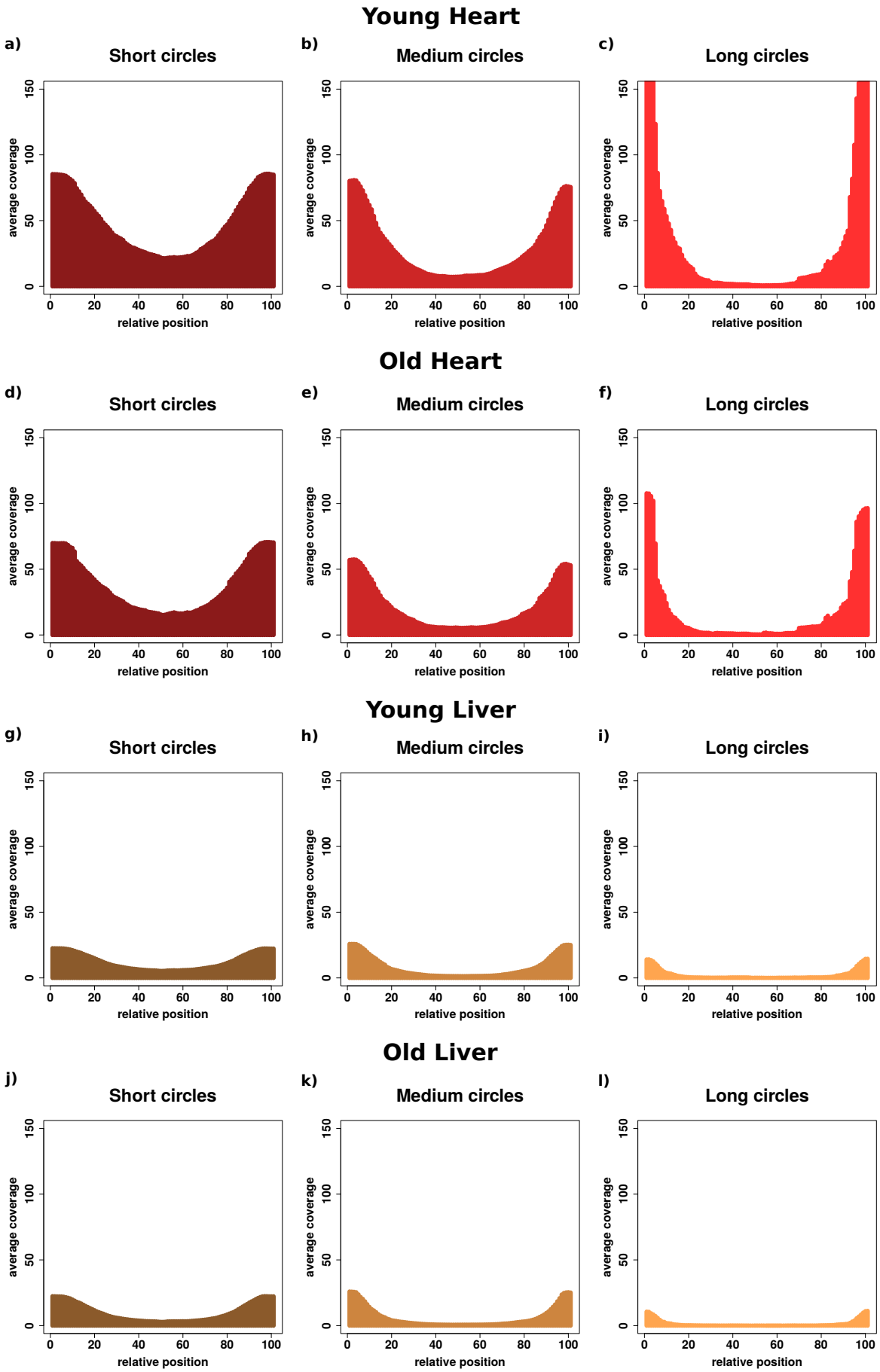

Figure 3.10: Coverage profiling results for each sample

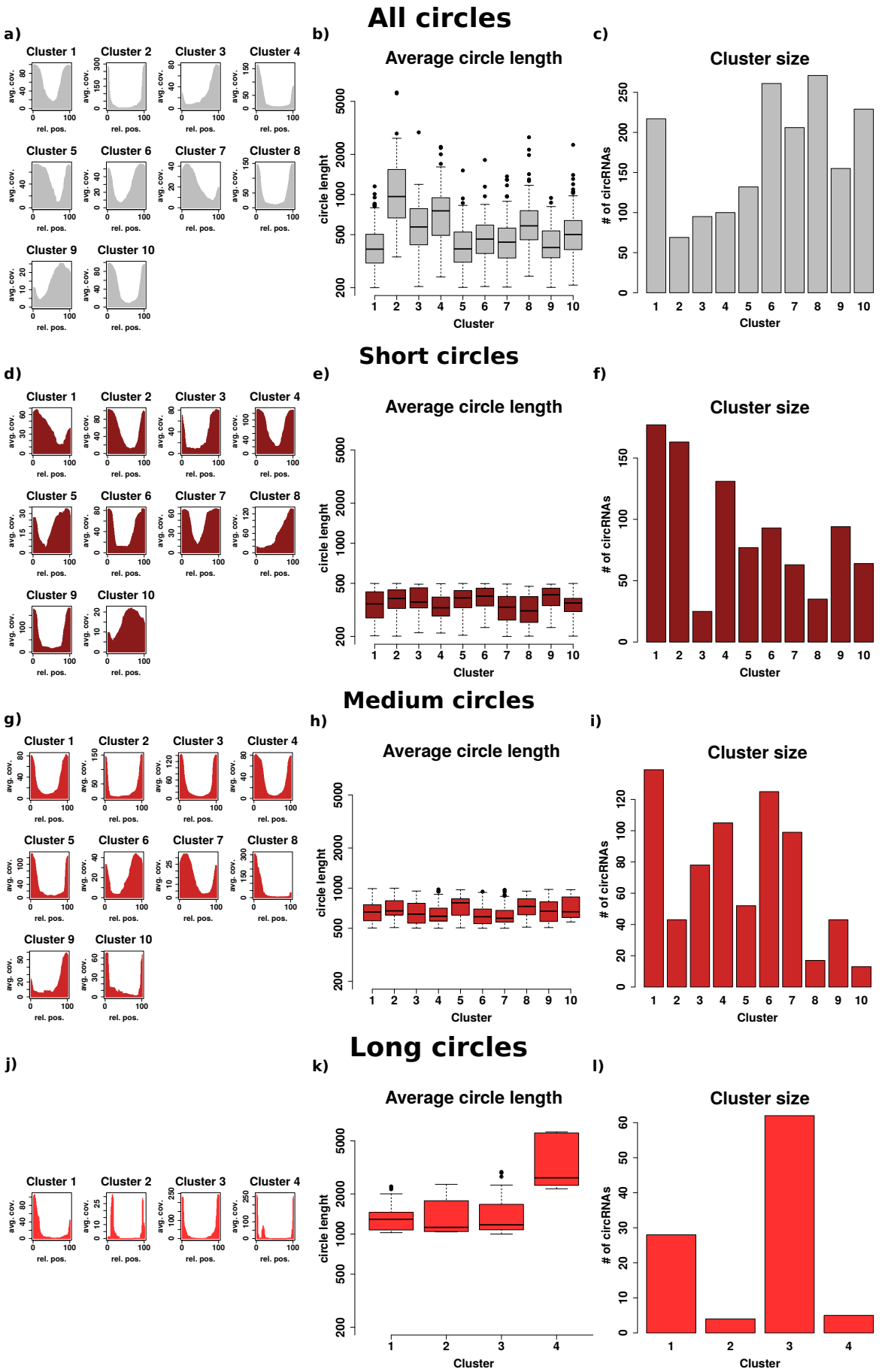

Figure 3.11: Example of the clustering by coverage profiles (Young Heart)

3.4 Conclusions and outlook

In summary, FUCHS is a new pipeline to summarize several aspects of the circRNA landscape and library characteristics. Summarizing circRNAs by host genes revealed that 40 % of host genes expressed more than one circular isoform. Analysing the proportion of double-breakpoint fragments unfolded a strong negative correlation between the median proportion of double-breakpoint fragments and the median circle length. Coverage profiling and clustering by coverage profiles informed about the characteristic distribution of reads around the circular RNA if only back-splice anchored reads were considered.

As shown in Figure 3.6, double-breakpoint fragments can originate from short overlapping reads or very long reads, which capture the whole circle and span the chimeric junction twice. I believe that with the correct library preparation double-breakpoint fragments may be used to indirectly validate circularization. First, the cDNA library would need to be synthesized before fragmentation of the library to allow for rolling circles. Second, a size selection would be necessary in order to sequence only fragments that are longer than twice the desired read length to avoid that reads overlap. Provided these library preparation steps, double-breakpoint fragments may only arise from long fragments, rolling around the circRNA at least twice.

Initially, I hypothesized that a coverage profile deviating from the norm would indicate another origin for the back-splice junction different from circularization such as trans-splicing or genomic rearrangement. However, the more likely reason for such a coverage profile is that the exon-intron structure of the circRNA did not match the annotated gene model as the coverage was only observed for annotated exons.

My conclusion is that FUCHS summarizes the circRNA landscape well but the detection of alternative splicing within the same back-splice junction is limited to annotated exons. In the next chapter, I will describe an additional method that attempts to overcome FUCHS's limitations to detect novel spliced isoforms by reconstructing the exon intron structure based on intron signals from back-splice junction anchored reads.

Chapter 4

De novo circle structure reconstruction based on intron signals – FUCHS*denovo*

4.1 Introduction

As described in the previous chapter, identifying inner circle structures based on annotation has the disadvantage that only annotated transcripts can be identified. The individual coverage profiles of circRNAs indicated that annotated gene model were not always suitable for representing the circle structures. Furthermore, alternative splicing events were underestimated as only annotated exons could be identified as skipped. For this reason I developed an additional program called FUCHS*denovo*. FUCHS*denovo* loads the separated circRNA BAM files generated using FUCHS and extracts the linear intron signals of back-splice junction anchored read pairs. First, the introns are connected into an intron chain. Second, exons are assumed to be in between two introns and then refined using the circRNA's coverage track. For circRNAs which are not fully covered, the exon-intron structure is inferred based on the provided annotation. This chapter will describe these steps in detail. To exemplify the functionality and results gained by running FUCHS*denovo* I will run the program on the previously mentioned murine heart and liver samples. Lastly, I will compare the results of FUCHS*denovo* and CIRI-AS to evaluate their performances.

4.2 Methods

Connect introns

In standard SAM or BAM files introns are indicated as `N` in the CIGAR string and the exact location of the intron start and end are saved in an extra field, `jI:B:i,position,position[,position,position]`. The breakpoint positions always come in pairs. FUCHS*denovo* uses `pysam` to parse the intron positions and saves them into a dictionary with the genomic coordinates as key and the reads spanning the respective intron as value. Once all reads are examined and all introns are noted, the introns are sorted by their coordinates.

Afterwards, the intron chain reconstruction phase starts. An empty dictionary is created to save all transcripts. This dictionary is initialized with one transcript $t[0] : [(Chr_{circle}, Start_{circle} - 1, Start_{circle})]$ to denote the start of all transcripts. Then, using a for loop to iterate through all transcripts trying to add the current intron to each transcript. The intron is chained to an existing transcript if the start of the current intron is bigger than the end of the transcript's last intron. If the current intron could not be chained to any existing transcript, a copy of all transcripts without their last intron is generated and the current intron is chained to suitable transcripts. Once all introns are connected, the circle end is added to all transcripts as $[(Chr_{circle}, End_{circle} + 1, End_{circle} + 2)]$ to close all transcripts (see Figure 4.1a).

Infer exons

Exons are inferred using the transcript dictionary with all introns connected as transcripts. Iterating through all transcripts, exons are assumed to exist in between two neighboring introns such that exon coordinates are: $[(Chr_{circle}, End_{intron[i]}, Start_{intron[i+1]} - 1)]$. Exons are written to the transcript dictionary, into the slot 'exons', with their coordinates as key and their average coverage as value. The average coverage is calculated based on the coverage track obtained with `pybedtools.coverage()`.

Creating exons in this naive way could result in long, false exons as indicated by the exon in red in Figure 4.1b. For example, if there were regions where no read aligned to, then there could not be an

intron signal which can be use for chaining. Therefore, the split coverage track is used to split and cut any exon with zeros in the coverage track as shown in Figure 4.1b.

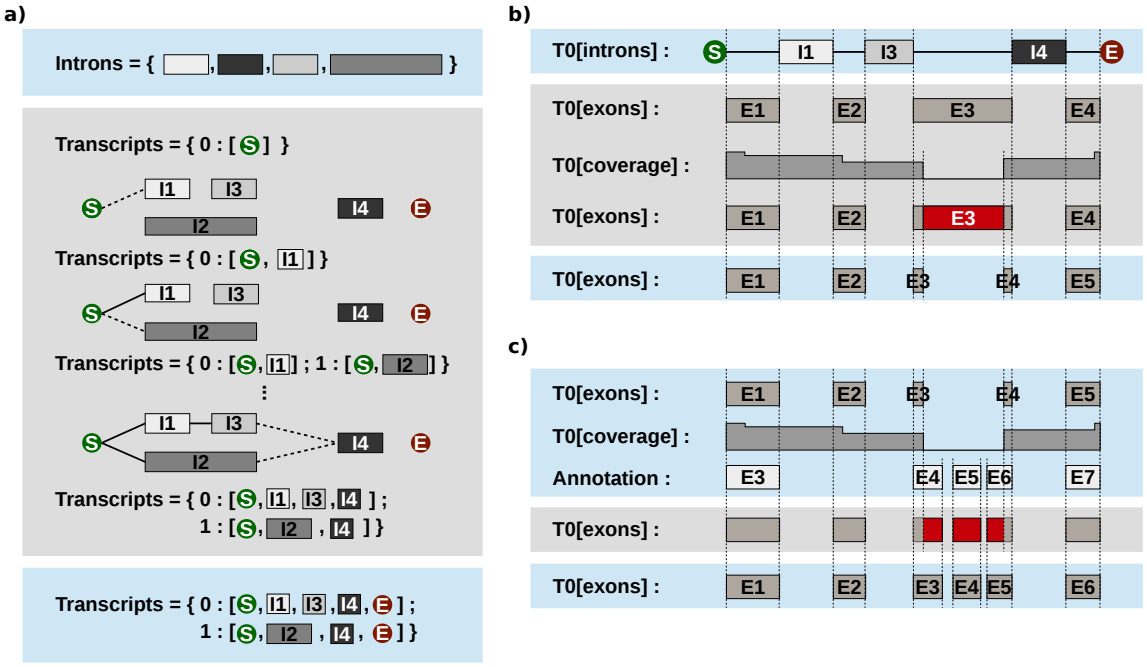

Figure 4.1: FUCHS*denovo*. **a)** Algorithm used to connect introns into transcripts. Light blue boxes represent the input and output of the method. Gray box represents the method. S and E represent the start and end coordinates of the circRNA. I1-I4 represent the forward-splice signals of back-splice junction anchored reads. The pseudo code represents Python based dictionaries and lists. **b)** Exon infering. The representation is the same as in the previous picture. E1-E4 represent exons. The red box labelled E3 represents a region of the circRNA which is not supported by reads. The horizontal dashed lines are only for orientation. **c)** Using annotated gene models to infer the intron exon structure of unsupported regions.

Infer structure in unsupported regions

If an annotation file was provided, FUCHS*denovo* infers the exon-intron structure in unsupported regions based on annotated exons using `pybedtools.intersect()`. Any identified exon is added to all transcripts with an average coverage of zero. This could result in overlapping exons if there are multiple annotated transcripts from different gene models. Therefore, exons are merged if they overlapped or if they were direct neighbors. Exons are considered to be direct neighbors if $\texttt{End}_{\texttt{exon[i]}} + 1 == \texttt{Start}_{\texttt{exon[i+1]}}$. The joining of exons stops once no exon overlaps another exon resulting in a gene model as exemplified in Figure 4.1c.

Handling of circRNAs without intron signals

There are two ways in which circRNAs may not have intron signals. First, the circRNA could be comprised of only one exon. Thus, there are no introns present. Second, it could be a long circRNA and only the first and last exons are supported by reads. Thus, no intron spanning reads are present in the circle BAM file. To report these circRNAs as accurately as possible, FUCHS*denovo* evaluates the coverage track of these circRNAs. If there are no zeros in the coverage track, the circRNA is written as a single exon circRNA. If there is a break in the coverage track, i.e. a continuous sequence of zeros, the structure is inferred as described in the previous subsection.

Write gene models

All reconstructed transcripts are written into two BED files. One exon-wise table in BED6 format (see Table 4.1 for more details) and one transcript-wise table in BED12 format (see Table 4.2 for more details).

Table 4.1: BED6

Chr	Start	End	Name	Score	Strand
10	128207298	128207390	10:128207298-128208667|0|0,1	17	.
10	128207538	128207622	10:128207298-128208667|1|0	6	.
10	128207560	128207622	10:128207298-128208667|2|1	6	.
10	128208590	128208667	10:128207298-128208667|3|0,1	18	.
10	21299178	21299243	10:21299178-21307779|0|0	4	.
10	21304529	21304643	10:21299178-21307779|1|0	4	.
10	21307584	21307779	10:21299178-21307779|2|0	5	.
10	123163593	123163654	10:123163593-123181800|0|0	2	.
10	123168226	123168372	10:123163593-123181800|1|0	0	.
10	123171099	123171226	10:123163593-123181800|2|0	0	.
10	123181672	123181800	10:123163593-123181800|3|0	1	.

Table 4.2: BED12

Chr	Start ThickEnd	End Color	Name NumberOfExons	Score ExonLength	Strand RelativeExonStart	ThickStart
10	128207298 128208667	128208667 255,0,0	10:128207298-128208667|0|1.0 3	4 92,84,77	. 0,240,1292	128207298
10	128207298 128208667	128208667 255,0,0	10:128207298-128208667|1|1.0 3	5 92,62,77	. 0,262,1292	128207298
10	21299178 21307779	21307779 255,0,0	10:21299178-21307779|0|1.0 3	4 65,114,195	. 0,5351,8406	21299178
10	123163593 123181800	123181800 255,0,0	10:123163593-123181800|0|0.263 4	1 61,146,127,128	. 0,4633,7506,18079	123163593

4.3 Results and discussion

De novo reconstruction reveals more internal alternative splicing

With FUCHS*denovo*, the number of host genes harboring circRNAs with different internal structures doubled to 8-10 % (compare Figure 3.7 and 4.2). This is concordant with previous findings presented in [46]. CIRI-AS predicted 5-10 % of all circRNAs to be alternatively spliced and 12-20 % of highly abundant ($\geq$ 20 BSJ reads) circRNAs to be alternatively spliced. Classifying these internal structures into four categories (AEU: alternative exon usage; A5SS: alternative 5' splice site; A3SS: alternative 3' splice site; and IR: intron retention) showed that for all samples the first three events were equally common, with 32 % (+/- 5 %) on average. Intron retention is the least observed alternative splicing event.

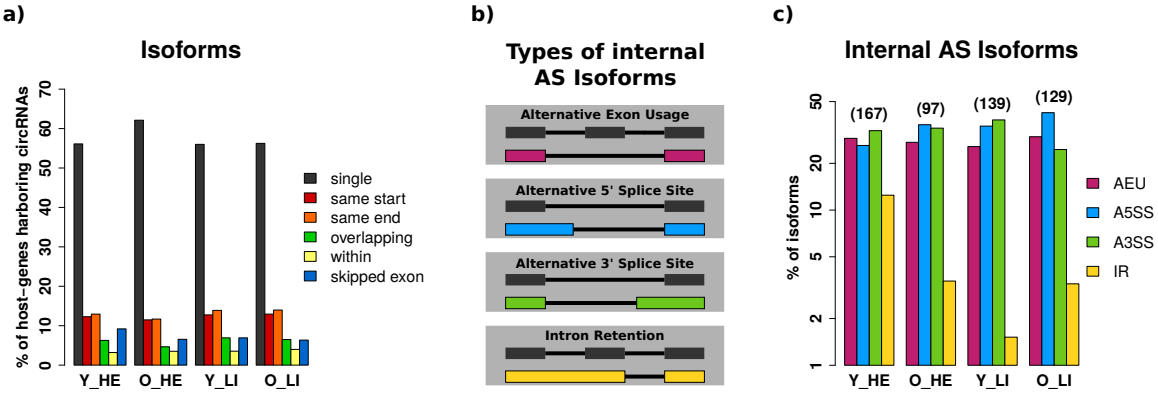

Figure 4.2: Host gene summary FUCHS*denovo*. **a)** Same data as in Figure 3.7 adjusted for the number of alternatively spliced circRNAs identified with FUCHS*denovo*. **b)** Four types of how a circRNA can be alternatively spliced. A combination of these splicing events is possible as well. **c)** Distribution of different alternative splicing events across the samples. The colors match the colors in **b.**

Alternative spliced isoforms are tissue specific

Figure 4.3 illustrates an alternative exon usage within the circRNA 5:135121506-135124017 of the transcription factor MLXIPL. The circRNA consists of five annotated exons. Additionally, there was one isoform skipping the fourth exon. While one isoform was common among all samples, the other isoform was not annotated and was only present in the liver samples but not in the heart samples. This indicates, that alternatively spliced circRNAs are tissue specific and not random miss-splicings or

miss-alignments. This was supported by the numbers of alternatively spliced circRNAs present in both samples of one tissue. The number of tissue-specific AS circRNAs were higher than the number of AS circRNAs present in all samples (see Figure 4.3).

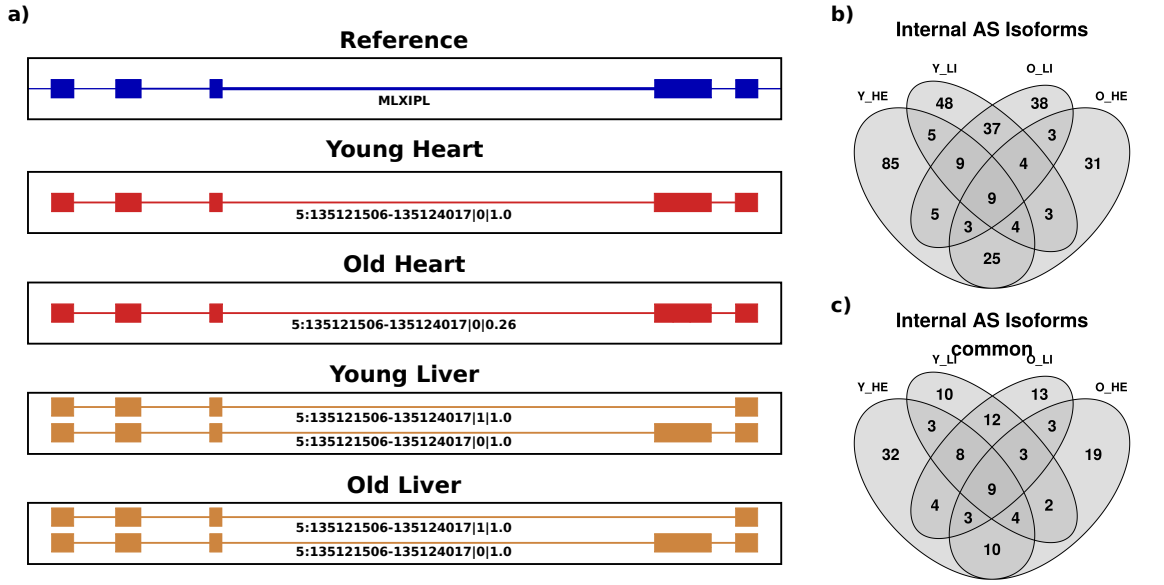

Figure 4.3: Internal circRNA structures. a) Exon-intron chains of two alternatively spliced circRNAs in the MLXIPL locus. Overlapping circRNA annotations were removed to enhance visibility of the alternative exon usage in circRNA 5:135121506-135124017. **b)** Venn diagram of the alternatively spliced isoforms present in one, two, three, or four samples. **c)** Venn diagram is a subset of the Venn diagram shown in **b**. Alternative splicing was only detected in circRNAs shared among all samples. Although circRNAs are present in all samples they are not always alternatively spliced in all samples.

Gain of more than 15% new information over previous approach

I examined the amount of information gained in order to show that reconstructing the exon-intron chain using intron signals first, and completing the circle structure using annotated features second, has an advantage over defining the circle structure using only annotated features. Exons were classified into four categories. If the circularized exons (circExon) were matching exactly to annotated exons (+/- 5 BP), they were classified as annotated exons. If circExons were not overlapping annotated exons, they were classified as completely new exons. If circExons were longer than the annotated exon, they were classified as partially new exons. If circExons were transcribed in the anti sense direction, they were classified as anti-sense exons. Figure 4.4a summarizes all categories as a graphical representation.

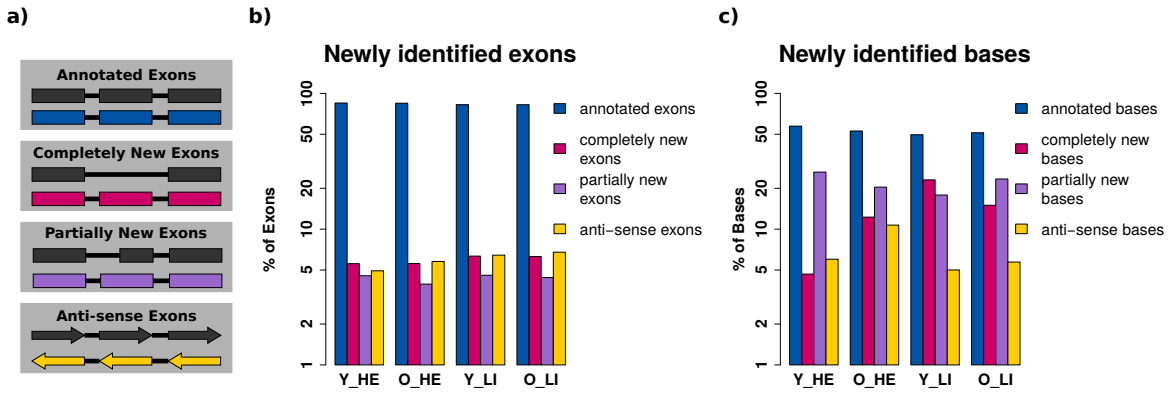

Figure 4.4: Gain of information over previous approach. a) Four classes that circRNA exons were classified as to compare the inferred structure to annotated gene models. **b)** Distribution of classes from **a**. The distrubtion is is equal over all samples with 85 % of exons matching fully to annotated exons and approximately 5 % of completely new, partially new, or anti-sense exons. **c)** Gain of information on a base level. Only not annotated bases were counted according to the class of their respective exon.

Evaluating the gain on an exon level revealed an even distribution of all categories across all samples with $\sim$ 85 % of exons matching annotated features and $\sim$ 5% of all other categories (see Figure 4.4b). When the resolution was increased to investigate how many bases this gain of information corresponded to the proportion of bases of exons exactly matching with annotated features decreased by 30 %. The proportion of new bases of partially annotated exons increased to 20 % on average. This is explained by alternative 5' and 3' splice site usage. The proportion of bases of completely new exons was highest in young liver and lowest in young heart. Considering that the proportion of new exons was equal in both samples it indicated that these new exons were shorter in heart than in liver (compare Figure 4.4b and c). The gain of 15 % on exon level and $>$ 40 % on base level shows that to study the function of circRNAs in silico, the *de novo* reconstruction of circRNAs is crucial in order to reliably detect miRNA seeds or RBP motifs in circularized sequences.

4.4 Comparison to CIRI-AS

Running CIRI-AS

CIRI-AS is the only other program that is able to detect alternatively spliced isoforms in circRNAs using the same raw data as FUCHS*denovo*. To run CIRI-AS one needs to detect circRNAs using CIRI first. Since its publication, CIRI was improved to use multi threading and to process reads of different length. However, CIRI-AS is neither able to use multi threading nor able to process reads of different length. Therefore, all reads were trimmed and filtered to 249 BP resulting in a loss of ~ 65 % of base pairs. The following commands were used to detect alternative splicing in both heart and liver samples using CIRI-AS:

```
# Trimming
flexbar -n 4 -r [sample]_R1.fastq.gz -p [sample]_R2.fastq.gz
             -t trimmed/[sample] -f sanger -u 50 -k 249 -m 249 -z GZ

# Aligning
bwa mem -P -T 19 -t 18 -B 3 bwa/GRCm38.79 trimmed/[sample]_1.fastq.gz
        trimmed/[sample]_2.fastq.gz > [sample].sam

# Detecting circRNAs
perl CIRI_v2.0.5.pl -I [sample].sam -O CIRI/[sample].ciri
                    -F GRCm38.dna.toplevel.fa -A GRCm38.79.gtf
                    -G CIRI/[sample].ciri.log -high -U 3 -T 8

# Detecting alternative splicing in circRNAs
perl CIRI_v2.0.5.pl -S [sample].sam -C CIRI/[sample].ciri
                    -O CIRI/[sample].ciri_as. -F GRCm38.dna.toplevel.fa
                    -A GRCm38.79.gtf -G CIRI/[sample].ciri_as.log
```

Evaluation parameters

Due to the lack of gold standards or positive controls, I compared FUCHS*denovo* and CIRI-AS only quantitatively by the circRNAs each programs detected as alternatively spliced. First, I compared the overlap of circRNAs used as input for FUCHS*denovo* and CIRI-AS, since both detection programs are only able to identify AS in previously detected circRNAs. Sample-wise four way Venn diagrams allow to evaluate how many circRNAs were identified as alternatively spliced by both programs, how many alternatively spliced circRNAs were overlooked by each program although the circRNA was detected, and how many alternatively spliced circRNAs were lost because of the difference in circRNA detection between DCC and CIRI.
Furthermore, I evaluated the runtime and the memory consumption of FUCHS*denovo* and CIRI-AS on equal computing environments.

FUCHS*denovo* and CIRI perform similarly well

First, comparing DCC and CIRI shows that half of the circRNAs detected by DCC were also detected by CIRI while two thirds of circRNAs detected by CIRI were detected by DCC (Figure 4.5a and b).
Of the circRNAs used to detect alternative splicing CIRI-AS identified a similar proportion of alternatively spliced circRNAs in all samples ($\sim$ 8 %). Nevertheless, since CIRI detected fewer circRNAs than DCC, the total number of alternatively spliced circRNAs was larger with FUCHS*denovo* than CIRI-AS (avg. 133 vs. avg. 101).
Examining the concordance of alternatively spliced circRNAs per sample showed that FUCHS*denovo* and CIRI-AS only agree on less than half of the alternatively spliced circRNAs. CIRI-AS failed to find the majority of circRNAs although CIRI detected the respective circRNA. However, without a positive control it remains elusive if these circRNAs are false positives identified by FUCHS*denovo* or false negatives not identified by CIRI-AS.
Evaluating the runtime and memory consumption showed that FUCHS*denovo* was faster on all evaluated samples but CIRI-AS was slightly more memory efficient. Because FUCHS*denovo* is able to use multi threading the runtime and memory demand can be tailored to the users requirements, i.e. trading off runtime for less memory.

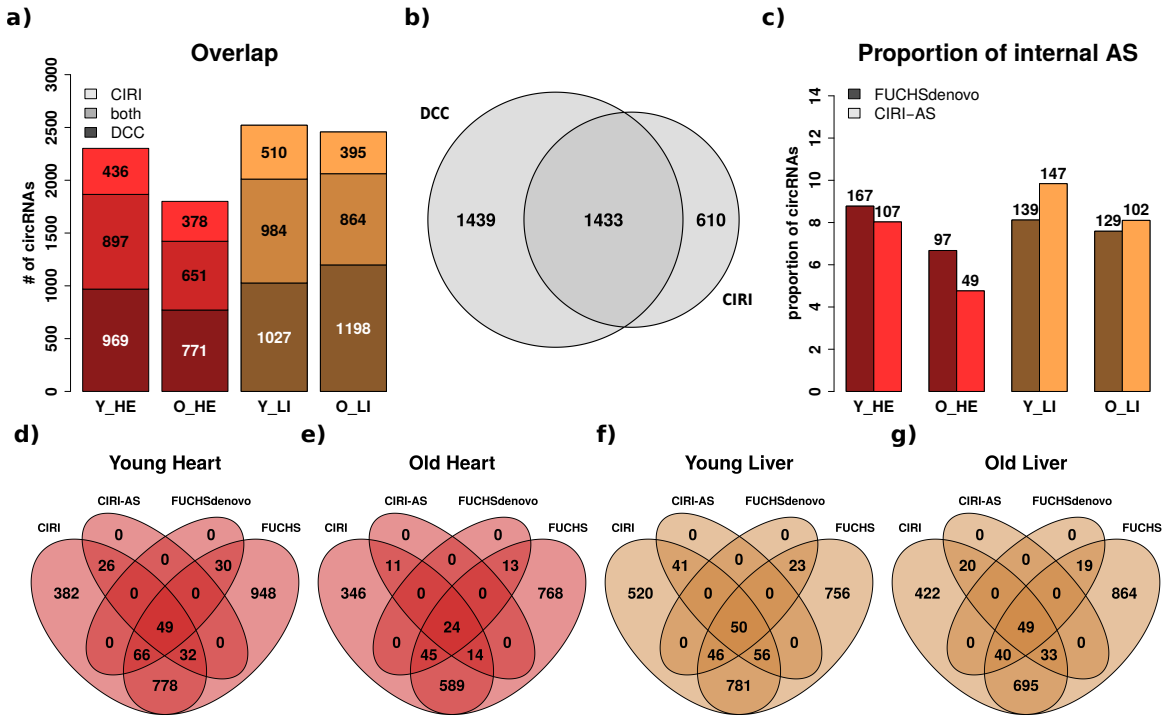

Figure 4.5: Comparison of FUCHS*denovo* and CIRI-AS. a) Stacked bar plot representing the number of circRNAs detected by DCC, CIRI, or both detection programs in each sample. **b)** Venn diagram of the overall overlap of circRNAs detected by DCC and CIRI. **c)** Overview of the estimated proportion of circRNAs FUCHS*denovo* or CIRI-AS predict as alternatively spliced. The number above each bar represents the total number of alternatively spliced circRNAs. Both programs predict the proportion of alternatively spliced circRNAs in a similar range for each sample. **d-g)** Venn diagrams summarizing the overlap of circRNAs, and alternatively spliced circRNAs for each sample. The overlap was performed on the circRNA coordinates. The outer two circles represent all circRNAs while the inner two circles represent the circRNAs with alternative splicing.

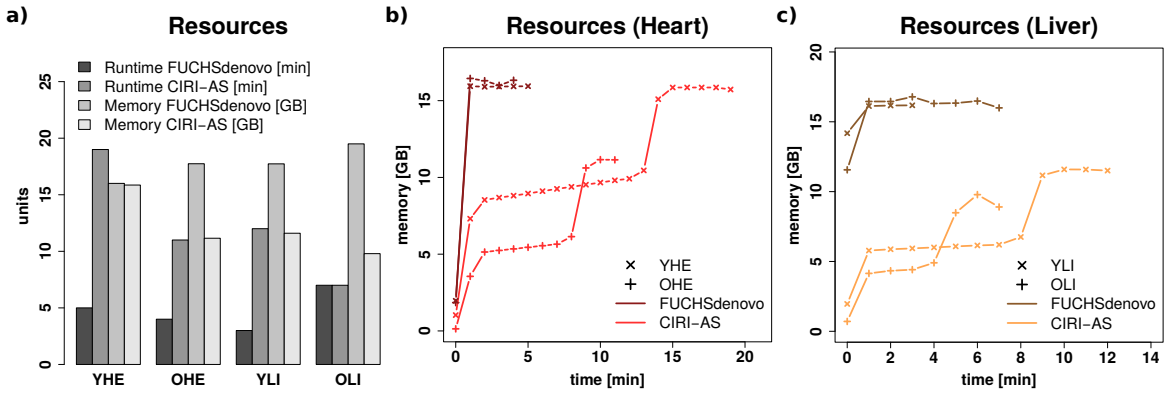

Figure 4.6: Runtime and memory consumption. a) Maximum memory and time that was consumed by each program for each sample. Both programs ran in the same computing environment. **b-c)** Memory consumption over time. While FUCHS*denovo* reaches its maximum memory within the first minute CIRI-AS memory consumption increases in two steps.

4.5 Conclusions

FUCHSdenovo is superior to the previous get_coverage_profile.py and detect_skipped_exons.py. It identifies twice as many alternative splicing events in circRNAs sharing the same back-splice junction. Evaluating the gain of information when running FUCHS*denovo* revealed a gain of 15 % on exon level and $>$ 40 % on base pair level over using annotated gene models to define the inner structure of circRNAs.

Classifying these alternative splicing events into different categories uncovered that AEU, A5SS, and A3SS were equally common. Intron retention is the least abundant alternatively spliced isoform. These results differ from previous reports where AEU was the most common AS isoform [46].

Comparing FUCHS*denovo* and CIRI-AS showed that both programs predicted an equal proportion of circRNAs to be alternatively spliced. On one hand, FUCHS*denovo* detected more circRNAs as alternatively spliced in CIRI-detected circRNAs. On the other hand, CIRI-AS predicted less circRNAs as alternatively spliced in DCC-detected circRNAs. Furthermore, FUCHS*denovo* is faster due to its ability to distribute the reconstruction of different circRNAs onto multiple threads while being comparably memory efficient as CIRI-AS.

Because FUCHS*denovo* reports the circle structure as BED6 and BED12 files, a feature lacking in CIRI-AS, the output of FUCHS*denovo* can directly be used to view the circle structures in a genome browser or to obtain the FASTA sequence to perform further computational analysis. In the following chapter I will describe a selection of these computational analyses that aim to decipher the function and biogenesis of circRNAs.

Chapter 5

Downstream analyses based on reconstructed circRNA structures

5.1 Introduction

Running `FUCHS` summarizes circRNAs by their host genes. Running `FUCHS`*denovo* additionally yields a more accurate circle structure than could be obtained using only annotated gene models. In the following chapter I will highlight possible computational downstream analyses that can be performed using the output of `FUCHS` and `FUCHS`*denovo*.

5.2 Methods

FASTA file generation

As previously discussed, FASTA files can be generated from BED files using `bedtools getfasta`. If a BED12 file is provided, `bedtools` is able to extract and concatenate the blocks of sequences into one transcript using the `-split` flag. Using the `-s` flag ensures that sequences are obtained with respect to their strand, i.e. for features on the reverse strand the reverse complement will be extracted. The FASTA sequences were generated with the following command:

```
bedtools getfasta -fi GRCm38.dna.toplevel.fa
                  -bed [sample]_exon_chain_inferred_12.strand_specific.bed
                  -fo [sample].fa -split -s
```

Host transcripts were filtered from an annotation file and transformed from BED6 to BED12 files in order to use the same `bedtools getfasta` parameters. Host transcripts and circRNAs were rearranged into one FASTA file per host gene containing all sequences of circRNAs and host transcripts belonging to that gene.

RBP motif and miRNA seed searches

To investigate whether any circRNAs could be targets of miRNA or RBPs, I performed a motif search on the previously described FASTA files using `FIMO` [49], a program integrated in the `MEME` suite[1]. In order to exploit the server's computing power, whole motif databases were scanned against gene-wise FASTA files containing circRNA and host transcript sequences, which were associated with the same gene as opposed to one FASTA file containing all sequences. Known miRNA seeds were downloaded from TargetScan[2] and transformed into a `MEME`-like motif database. For known RBP motifs, the `MEME` database 'uniprobe_mouse.meme'[3] was downloaded from the `MEME` suite. `FIMO` version 4.11.2 was used to compare these databases to the target sequences using the following command:

```
fimo -v 1 --thresh 0.0001 --norc -- oc [sample]/[gene]/
     [database].meme [sample]/[gene].fa
```

The identified seed hits were merged into one table. p-values were corrected for multiple-hypothesis testing using the R function `p.adjust(method = 'fdr')`. Hits with an FDR > 0.01 were removed from further analysis.

To test whether circRNAs were more densely populated with miRNA seeds or RBP motifs, the number of hits per transcript was divided by the length of the corresponding FASTA sequence.

To identify meaningful circRNA-miRNA or circRNA-RBP pairs, the number of seeds was summarized by circRNA and miRNA, RBP respectively. p-values for circRNA-miRNA or circRNA-RBP pairs

[1] http://meme-suite.org/
[2] http://www.targetscan.org/mmu_71/mmu_71_data_download/miR_Family_Info.txt.zip
[3] http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.15.tgz

were calculated by permuting each motif in both databases 1000 times, as previously described by Guo *et al.* [50]. RBP motifs were complex enough to guarantee 1000 unique permutations of each motif; however, the majority of miRNA seed motifs gave rise to fewer than 1000 unique motifs. Therefore, the maximum number of unique permutations was used to calculate the *p*-value for each circRNA-miRNA pair.

GO enrichment

To gain insights into which functions or processes certain circRNAs could play a role in, GO enrichment analyses were performed on different sets of host genes (Table 5.1). The enrichment was performed using the web interface of `GOrilla`[4] choosing *M. musculus* as organism and providing a target and background gene list. All three GO categories were analysed setting the *p*-value threshold to report enriched GO terms to 0.001.

Table 5.1: GO enrichment input sets

Target set	Background set	Question
Host genes expressing more than one circRNA	All host genes	Are genes that express multiple circRNAs important for a specific process compared to all host genes?
Host genes of circRNAs with alternatively spliced isoforms	All host genes	What are possible biological reasons for alternative splicing?
Host genes of circRNAs enriched for miRNA seeds	All expressed genes	What are possible consequences if circRNAs were to compete for miRNAs?
Host genes of circRNAs enriched for RBP motifs	All expressed genes	What are possible consequences if circRNAs were to compete for RBPs?

Motif search in flanking introns

It is known that flanking introns are an essential factor driving circularization. Here I wanted to investigate if there were any motifs differentially enriched between circRNAs of certain characteristics. I generated FASTA files of the flanking introns. The sequences comprised of 500, 2000, or 5000 BP upstream of the splice donor or downstream of the splice acceptor of each back-splice junction. The flanking introns of circRNA groups of interest (see Table 5.2) were used as the target set, while

[4]http://cbl-gorilla.cs.technion.ac.il/

those of other circRNAs were grouped into a background set. A differential-enrichment analysis was performed between the target and background set to ascertain if there were any motifs associated with the circRNAs of interest.

These enrichment analyses were performed using `DREME` [51], which compares the frequency of motif hits between two sets of sequences using Fisher's exact test:

```
dreme -v 1 -oc [set]_[size] -dna -p [set]_[size]_targets.fasta
      -n [set]_[size]_background.fasta -norc -e 0.05 -eps
```

Table 5.2: Overview over the sets of flanking introns that were compared using `DREME`

Name	Target	Background
Host-gene-wise	all circRNAs from host genes that express multiple circRNAs	all circRNAs from host genes that only express one circRNA
Splice-site-wise	only flanking introns of splice site which were used more than once	flanking introns of splice sites which were only used once
Isoform-wise	circRNAs that show alternative splicing	circRNAs that only have one exon-intron chain

5.3 Results and discussion

miRNA seed search

Comparing the total density of miRNA seeds in circRNAs to that in their host transcripts revealed that the overall seed density in the former was higher than in the latter (see Figure 5.1a). The seed density was higher in circRNAs expressed in liver than those expressed in cardiac cells, while the density of seeds in host transcripts did not differ in these tissues. This could indicate a possible general sponge function. Some circRNAs were not only more densely populated with miRNA seeds but harbored more seeds than their host gene when considering the average over all transcripts (see Figure 5.1b). This could be explained by alternative splice site or exon usage in circRNAs as described previously (Chapter 4.3). The histogram in Figure 5.1c shows that most circRNAs only harbored two to three miRNA seeds of a given miRNA. One notable exception was circRNA 15:3388648-3551685 containing 18 seeds for miRNA-29-3p. The circRNA originated from the growth hormone receptor

(GHR) locus. Nevertheless, I could not identify any circRNA as specifically enriched as CDR1-AS, which harbors 70 seeds for miR-7 [22, 23].

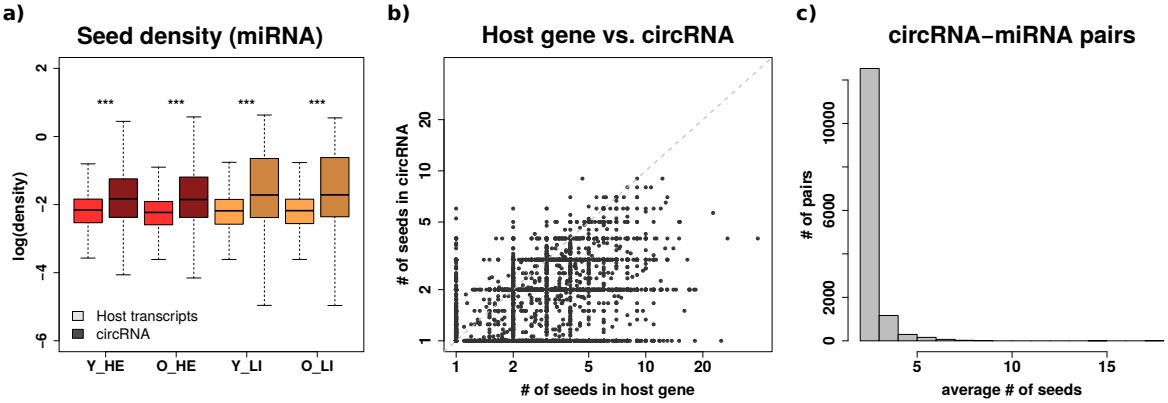

Figure 5.1: Results of the miRNA seed search. a) Overall density of miRNA seeds in either host transcripts or circRNAs for each sample. Only seeds with a q-value ≤ 0.01 were considered for this analysis. CircRNAs are significantly more densely populated with miRNA seeds than host transcripts ($p \leq$ 2.2e-16). **b)** Number of seeds in host genes (averaged over all transcripts) vs. the number of seeds in circRNAs. The gray dashed line serves as orientation (y = x). **c)** Overview of significant circRNA-miRNA pairs. p-values were calculated by permutation for each specific circRNA-miRNA pair. The histogram shows only significant pairs ($q \leq 0.05$).

RBP motif search

The results of the RBP motif search were similar to those of the miRNA seed search: the total motif density was significantly increased in circRNAs over their host transcripts (see Figure 5.2a), and the average number of motifs over all samples of a single RBP ranged from two to three. The strongest motif enrichment was observed in circRNA 16:17671545-17674658 of the mediator complex subunit 15 (MED15) gene, which contained nine binding sites for ASCL2 (see 5.2c).

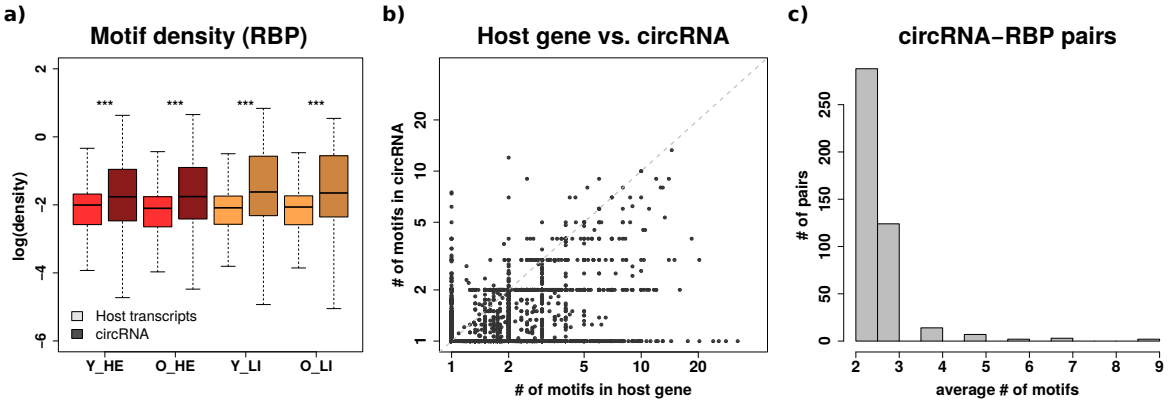

Figure 5.2: Results of the RBP motif search. a) Overall density of RBP motifs in either host transcripts or circRNAs for each sample. Only motifs with a q-value ≤ 0.01 were considered for this analysis. CircRNAs are significantly more densely populated with RBP motifs than host transcripts ($p \leq$ 2.2e-16). **b)** Number of motifs in host genes (averaged over all transcripts) vs. the number of motifs in circRNAs. The gray dashed line serves as orientation (y = x). **c)** Overview of significant circRNA-RBP pairs. p-values were calculated by permutation for each specific circRNA-RBP pair. The histogram shows only significant pairs ($q \leq 0.05$).

GO enrichment

In Chapter 2, GO enrichment was performed on circRNAs expressed in liver, heart, or both against all expressed genes, in order to investigate which GO terms are enriched in circRNA host genes. Here, I investigated which GO terms host genes expressing multiple circRNAs and host genes expressing alternatively spliced circRNAs were enriched for. Furthermore, I tested host genes of the top ten percent of miRNA seeds or RBP motifs enriched circRNAs to gain information of potential consequences in case circRNAs competed with their host genes for miRNA or RBP binding. Table 5.3 shows the results of all performed GO enrichments. The class of host genes bearing multiple circRNAs was enriched for various regulatory processes, including regulation of metabolic processes and microtuble cytoskeleton organizations. Host genes of alternatively spliced circRNAs were not enriched for specific GO terms.

miRNA seeds and RBP motifs were enriched in circRNAs whose host genes were involved in various regulatory and metabolic processes such as organic substance metabolic process and regulation of biological processes. These results indicate that circRNAs could buffer regulatory processes by competing for miRNA and RBP binding.

Table 5.3: GO enrichment

Enriched in	GO Term	FDR	E
Multi circRNA host genes (601 genes)	regulation of cellular process	2.84E-2	1.12
	regulation of macromolecule metabolic process	1.25E-1	1.17
	regulation of biological process	1.49E-1	1.10
	regulation of metabolic process	2.29E-1	1.14
	biological regulation	2.05E-1	1.09
	regulation of microtubule cytoskeleton organization	2.71E-1	2.13
	regulation of cellular metabolic process	2.61E-1	1.14
	positive regulation of molecular function	2.33E-1	1.34
	regulation of primary metabolic process	2.91E-1	1.14
	animal organ morphogenesis	2.66E-1	1.62
	regulation of cellular macromolecule biosynthetic process	2.82E-1	1.19
	regulation of macromolecule biosynthetic process	3.83E-1	1.18
	negative regulation of MAPK cascade	3.79E-1	1.97
	plasma membrane organization	4.15E-1	1.87
Enriched in AS circRNAs (248 genes)			
Enriched in top 10 % miRNA seed enriched circRNA (396 genes)	cellular metabolic process	1.25E-17	1.64
	regulation of biological process	3.26E-11	1.31
	primary metabolic process	1.67E-16	1.60
	organic substance metabolic process	3.99E-16	1.57
	cellular component organization	7.60E-13	1.67
Enriched in top 10 % RBP motif enriched circRNA (47 genes)	regulation of RNA metabolic process	6.78E-2	2.70
	cellular component organization or biogenesis	2.96E-2	2.38
	developmental process	5.82E-2	2.13
	biological regulation	5.90E-2	1.54
	cellular process	1.21E-1	1.40

GO terms reported with $p < 0.001$, 1624/2828 background genes

Flanking introns

Performing a differential motif enrichment analysis of binding motifs in the flanking introns revealed several motifs enriched in different regions (0-500, 0-2000, and 0-5000 BP of the flanking intron) of circRNAs of interest (see Figure 5.3 host-gene-wise). A comparison of these motifs against known motifs using `TOMTOM` [52] revealed three interesting matches, CPEB1, HOX, and FOXO.

The motifs matching to the cytoplasmic polyadenylation element binding protein (CPEB1) were enriched in the flanking introns of circRNAs from host genes yielding multiple circRNAs (see Figure 5.3). CPEB1 has been shown to influence alternative processing of its target genes and has been indicated in chronic liver disease [53]. It could also be involved in alternative circRNA formation in liver.

The motifs matching to members of the homeotic gene family (HOX) were also enriched in the flanking introns of circRNAs from host genes yielding multiple circRNAs. HOX genes are transcription factors active during in development [54, 55]. It is possible that HOX is responsible for the transcription of circularized exons, thus making exon circularization more efficient. To my knowledge, neither HOX nor CPEB1 have been indicated to play a role in circle formation, yet.

Interestingly, the immediate surroundings of alternatively spliced circRNAs revealed a positive matche for FOXO3. This motif is also matching to FOXO1, FOXO6, and FOXO4. Forkhead box class O (FOXO) proteins are transcription factors regulating several developmental processes, including cell growth, proliferation and differentiation [56–58]. They have also been linked to longevity [59]. FOXO proteins have not previously been implicated in circularization, however, the circRNA of FOXO3 has been linked to cardiac senescence [60].

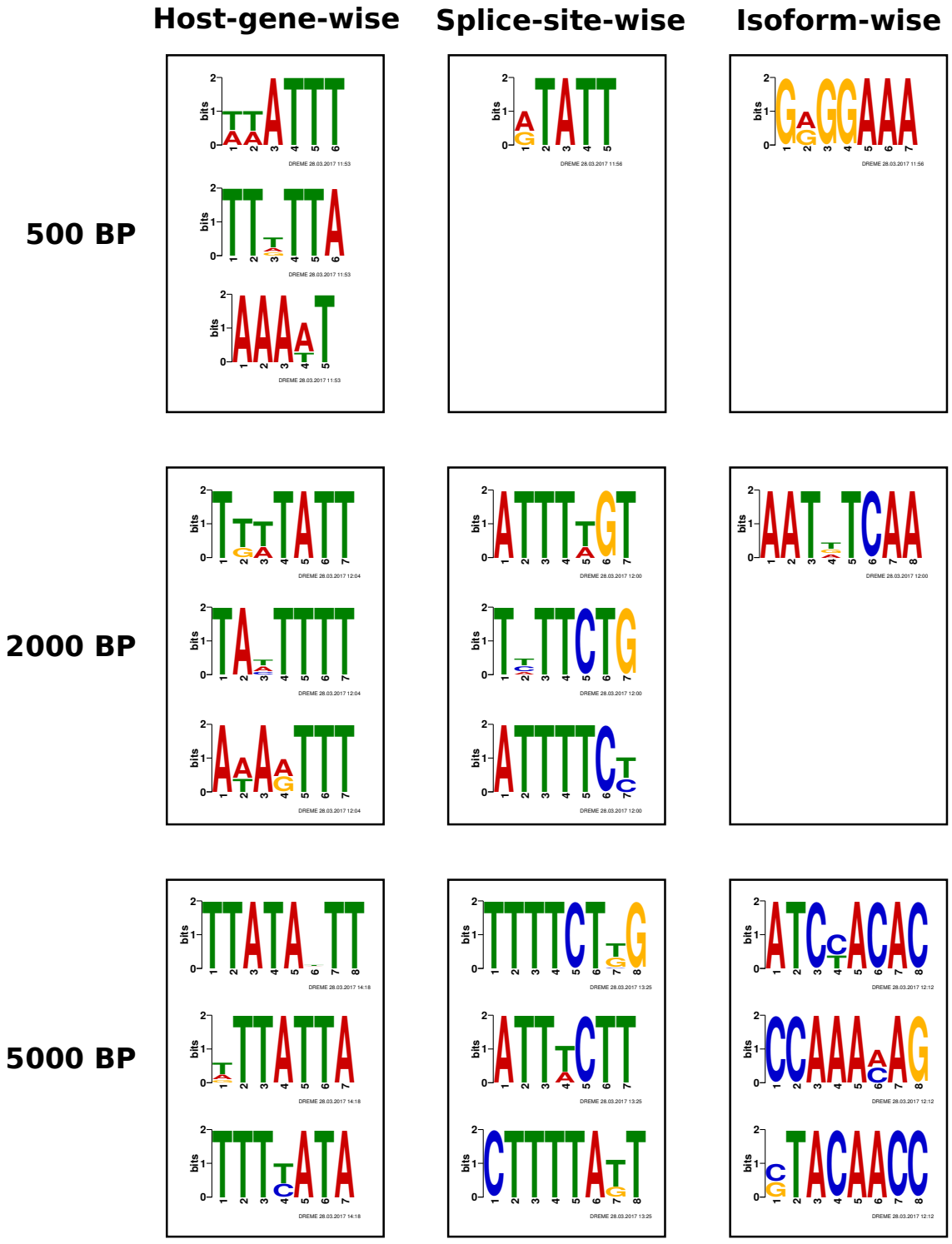

Figure 5.3: Differentially enriched motifs. Differentially enriched motifs in the flanking introns of different groups of circRNAs. Only the top three motifs for each analysis are shown here.

5.4 Conclusions

The analysis performed in this chapter demonstrated potential downstream analyses that can be performed based on the circRNA structures reconstructed by `FUCHS`*denovo*. The BED6 and BED12 files produced by `FUCHS`*denovo* can be used to acquire the FASTA sequences. Using `bedtools getfasta` to obtain the FASTA sequence either exon-wise with the BED6 file or the complete circle structure with the BED12 file. These sequences can then be used to find miRNA seeds and RBP motifs. Other sequence based analyses such as differential enrichment analysis of identified motifs are also possible.

The miRNA seed and RBP motif search showed that circRNAs are more densely populated with seeds/motifs than their full-length host transcripts. This suggest that circRNAs could constitute another layer of regulation by competing with their host genes for miRNA or RBP binding. However, I could not identify a circRNA with a clear sponging function like that of CDR1-AS [22, 23].

Using the information gained by running `FUCHS` to perform a GO enrichment on host genes expressing many circRNAs over host genes expressing only one circRNA revealed that these host genes are enriched in regulatory processes.

Performing differential motif-enrichment analysis between the flanking introns of circRNAs from multi- and single-circRNA host genes yielded two protein families, HOX and CPEB1, that could potentially be involved in circRNA formation. Furthermore, flanking introns of alternatively spliced circRNAs carry a FOXO motif. These protein families have not previously been linked to circle formation, but pose interesting targets of future investigations to help understand the circRNA biogenesis.

Chapter 6

Summary and conclusions

In my thesis, I benchmarked three state of the art circRNA detection programs and showed that DCC is the most accurate and fastest circRNA detection program. CIRI performed second best while KNIFE ranked third.

In chapter 3, I introduced my own tool to summarize and characterize circRNAs identified by DCC. My program is called FUCHS, short for **FU**ll **CH**aracterization of circular RNA using RNA-**S**equencing. FUCHS identifies host genes expressing several circRNAs, detects exon skipping of annotated exons, calculates the proportion of double breakpoint reads, and generates circle coverage profiles. The circle coverage profiles indicated that the annotated gene models were not always suited to represent the circle's exon-intron structure. Therefore, I extended FUCHS by a new module which I called FUCHS*denovo*. FUCHS*denovo* reconstructs the exon-intron structure; first, by chaining together linear intron signals of back-splice junction anchored read pairs. Second, if an annotation is provided, it infers the exon structure of unsupported regions based on annotated gene models.

Throughout the thesis I used a sample dataset of two murine hearts (young and old) and two murine livers (young and old) (see Table 2.3). The libraries were enriched for circular RNAs using *RNaseR* treatment and sequenced with 250 BP paired-end reads in order to fully capture as many circRNAs as possible. These samples were aligned using STAR. Their circRNAs were detected using DCC. Identified circRNAs were summarized and characterized using FUCHS and their circle structure was defined using FUCHS*denovo*. Because there was only one library per tissue and time point it was only possible to describe the circRNA landscape quantitatively. The analysis showed that heart samples

were less diverse but expressed circRNAs were more abundant than in liver. A GO enrichment indicated that genes expressing circRNAs were enriched for metabolic regulation. Comparing the length of circRNAs revealed that the circRNAs in heart were longer than in liver. The summary of circRNAs by host genes revealed that circRNAs from the same host gene often shared the same 5' or 3' splice site. Alternative splicing of circRNAs sharing the same back-splice junction occurred in 5-10 % of all circRNAs. Alternative 5' or 3' splice site usage and alternative exon usage were equally abundant while intron retention was the least prevalent (< 10 %) alternative splice isoform.

Using `bedtools` to obtain the FASTA sequences of the reconstructed circRNA structures allowed to perform sequence based analysis such as scanning for miRNA seeds or RBP motifs. CircRNAs in heart and liver were more densely populated with miRNA seeds and RBP motifs than their host genes. However, no outstanding circRNA – miRNA/circRNA – RBP pair such as CDR1-AS – miR-7 could be found. My results indicate that circRNAs provide an additional layer for the regulation of transcript expression as they are likely to compete with their host genes for miRNA and RBP binding. A differential motif analysis of the flanking introns revealed that CPEB1 and HOX could be involved in circRNA formation from host genes resulting in multiple circRNAs while FOXO proteins may be involved in the formation of alternatively spliced circRNAs.

Essentially, this work introduces a new tool, `FUCHS`, to study the circRNA landscape of curated organisms. The additional module, `FUCHS`*denovo*, provides the foundation for further sequence based computational analysis.

Chapter 7

Outlook: circRNAs in the African turqoise killifish

7.1 Introduction

CircRNAs in ageing

The function of most circRNAs remains elusive. Enuka *et al.* [6] stimulated MCF10A cells with the epidermal growth factor (EGF) to investigate if circRNAs are involved in highly dynamic signalling cascades. They observed the expression of mRNAs, miRNAs, and circRNAs before stimulation and up to four hours after stimulation. While the expression of many mRNAs as well as miRNAs changed dramatically, the expression of circRNAs remained stable during and after the stimulation. This result, together with the long half life of circRNAs, led Enuka *et al.* to the hypothesis that circRNAs are not involved in signal cascades of fast and dynamic processes, but rather long-term processes such as differentiation and cell ageing.

This hypothesis is supported by work from You *et al.* [48]. First, they identified brain as the tissue with the highest circRNA expression in mice and later followed the expression of circRNAs during the development of mouse hippocampi. In total, 224 circRNAs displayed a differential expression pattern during development (181 were up-regulated, whereas 43 were down-regulated). The genes encompassing these 181 up-regulated circRNAs were over-represented for pathways involved in

synaptic function. This led You *et al.* to hypothesise that circRNAs are involved in synaptic development and synaptic plasticity.

These studies imply that circRNAs could play a role during ageing. While Westholm *at el.* [34] showed that the overall expression of circRNAs in the nervous system of *D. melanogaster* increases during age, no study has investigated the function of circRNAs during ageing in vertebrate models, yet.

CircRNAs in the brain

Several studies comparing different tissues showed that brain samples were significantly enriched with circular RNAs, not only with respect to abundance, but also with respect to diversity [34, 48, 61]. One of the first studies that indicated an enrichment of circRNAs in the central nervous system (CNS) over other tissues was performed by Westholm *et al.* [34]. More than 100 libraries were sequenced from different developmental stages and tissues from fruit fly. A GO enrichment analysis showed that circular RNA expressing genes were enriched for development, signalling and neural function. More importantly, neural function was also increased during development before the CNS was formed. Westholm *et al.* observed an increase of overall circular expression in ageing flies, but did not speculate if this increase was due to accumulation or specific regulation and suggested, that although this increase of circular RNAs may be incidental it may still be of biological relevance.

A more recent study investigated the circRNA expression in porcine embryonic brain development as the first circRNA study on large mammals [62]. Contrary to the mouse and fly studies mentioned before, they found a steady increase of the relative amount of circRNAs until day sixty of embryonic development followed by a dramatic decrease from E60 to E80, flattening out until the day of birth (E115). The most enriched pathways of host genes expressing circular RNA of > 0.15 RPM at E60 are Wnt signalling, axon guidance, and TGF-beta signalling. Other observations including that circRNA flanking introns are longer correspond to previous studies.

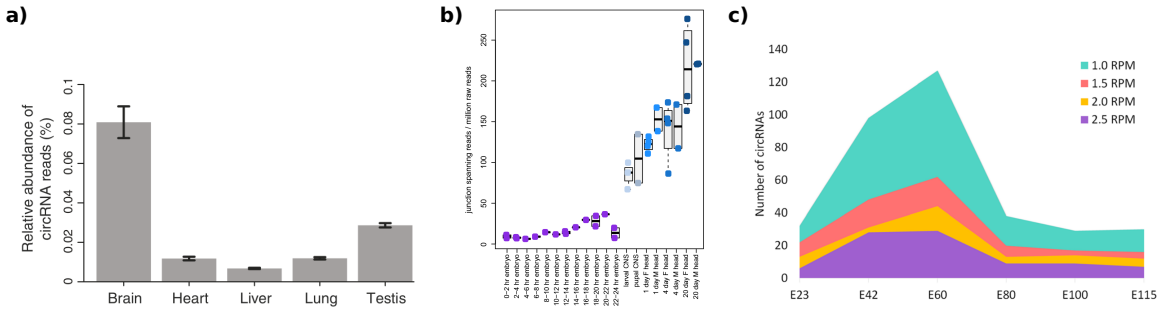

Figure 7.1: CircRNAs are highly enriched in brain samples. **a)** circRNA expression in different mouse tissues [48]. **b)** circRNA expression in fruit fly heads [34]. **c)** circRNA expression in porcine brain [62].

A new model organism for ageing

The traditional model organisms to study ageing are *C. elegans* and *D. melanogaster*; they have the advantages of a short life cycle, high number of offspring, and relatively low maintenance cost. Furthermore, numerous genetic tools and lines are available to study gene function during ageing. Other model organisms include *D. rerio* and *M. musculus*. They are more closely related to humans than *C. elegans* or *D. melanogaster*, but their life span is too long for experimental manipulations. Recently, a new model organism has been introduced to the ageing research. The African turquoise killifish has the advantage of being a vertebrate with a much shorter life span than mice and zebrafish (compare Fig. 7.2), while showing symptoms of ageing phenotypes also found in humans. These include loss of pigmentation, reduction of locomotor activity, and neurodegeneration [63, 64].

Using this new model, I investigated the circRNA landscape during the life of the African turquoise killifish, and established the over-expression of circRNA candidates in the fish to study the function of circRNAs *in vivo*.

Here I am presenting the results of twelve killifish samples from a pilot RNA-Seq.

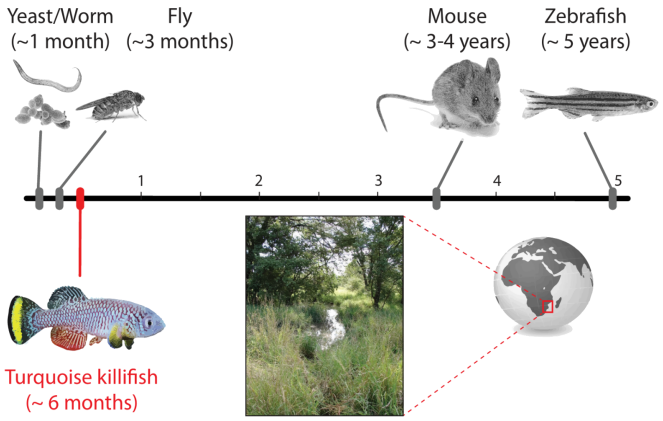

Figure 7.2: Model organisms for ageing research. The x-axis is in years. In ageing research short lived invertebrates or longer lived vertebrates are used for experiments due to the lack of a short lived vertebrate model. The African turquoise killifish fills this gap as a extremely short lived vertebrate. The bottom right map shows the natural habitat of the African turquoise killifish. This figure was taken from [65].

7.2 Methods

Data

Samples of young, adult and old fish were obtained (Fig. 7.3). Two replicates for each time point from brain and muscle were sequenced with 100 BP single-end reads of rRNA-depleted RNA-Seq libraries to test if circRNAs were expressed in the killifish (see Table 7.2 for details).

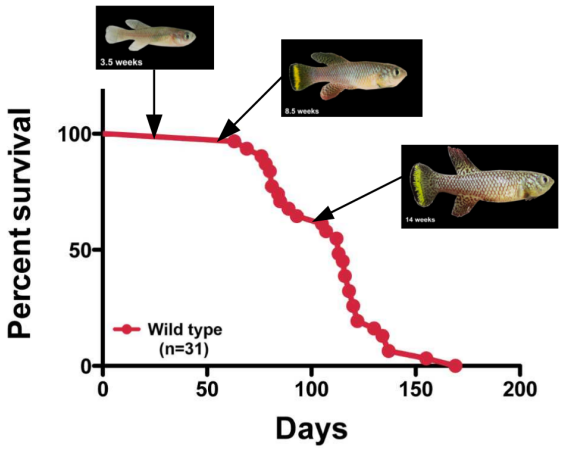

Figure 7.3: Sampling time points. Overview of the three sampling time points that were taken for this analysis. Fish were sampled at 3.5 weeks (before sexual maturation); 8.5 weeks (at the peak of sexual maturation); 14 weeks (median life span). Samples, pictures and lifespan curve were provided by Dr. Yumi Kim.

Table 7.1: Sample Information

Sample name	Tissue	Age [w]	Strain	Replicates	Library
Y_muscle_1-2	muscle	3.5	killifish (GRZ-AD)	2	100 BP SE ribo-
YA_muscle_1-2	muscle	8.5	killifish (GRZ-AD)	2	100 BP SE ribo-
O_muscle_1-2	muscle	14.0	killifish (GRZ-AD)	2	100 BP SE ribo-
Y_brain_1-2	brain	3.5	killifish (GRZ-AD)	2	100 BP SE ribo-
YA_brain_1-2	brain	8.5	killifish (GRZ-AD)	2	100 BP SE ribo-
O_brain_1-2	brain	14.0	killifish (GRZ-AD)	2	100 BP SE ribo-

Tissue-specific circRNA expression

To describe the circRNA landscape in different tissues, I calculated the diversity as the number of different circRNAs expressed in a given tissue. I used the R package `VennDiagram` to overlap the circRNA coordinates and to group the circRNAs into three groups: muscle-specific, brain-specific, and core circRNAs.

Conservation

To identify circRNAs that may play a role in human cells, I searched for conserved circRNAs. I downloaded the circRNA annotation tables of human and mice from circbase[1] [66] and overlapped the gene lists based on the names of genes expressing circRNAs in either human, mouse, or killifish.

GO enrichment

To obtain a first impression of the function genes expressing circRNAs in either brain or muscle I performed a GO enrichment analysis using the online platform `GOrilla` [44]. I mapped the killifish gene names to human gene names, and ran the enrichment analysis using a human reference database. For tissue specific circRNAs I used all the genes expressed in the given tissue as background list and the circRNA host genes as target list.

Temporal clustering

Besides the tissue specificity, I was interested in the temporal behaviour of the circRNA expression. I performed a clustering of the circRNAs on their expression patterns using the R package

[1] http://circbase.org/

`amap`'s kmeans algorithm with the following parameters: `Kmeans(ratio, centers = 8, method = 'correlation', iter.max = 20, nstart = 8)`. This way I was able to easily identify clusters of circRNAs showing an increase or decrease with age relative to their host gene.

Candidate validation

To validate back-splice junctions using qPCR, candidate circRNAs were selected based on their temporal expression (increasing with age) and conservation across species. Outward facing primers pairs were designed around the back-splice junction to quantify the circRNA expression. Additional primer pairs were designed from one side of the back-splice junction to the next exon on the linear transcript to quantify the host gene expression.

Candidate overexpression

Based on the temporal expression, conservation, validation by qPCR, and biological relevance candidate genes were selected for overexpression *in vivo*. Target exons were cloned into the pcDNA3.1(+) circRNA miniVector ([16], Fig. 7.4). This vector harbors specific Alu repeats on both ends of the customizable region to facilitate exon circularization. This vector was cloned into the Tol2-killifish vector [67] and injected during the one-cell stage of the fish embryo.

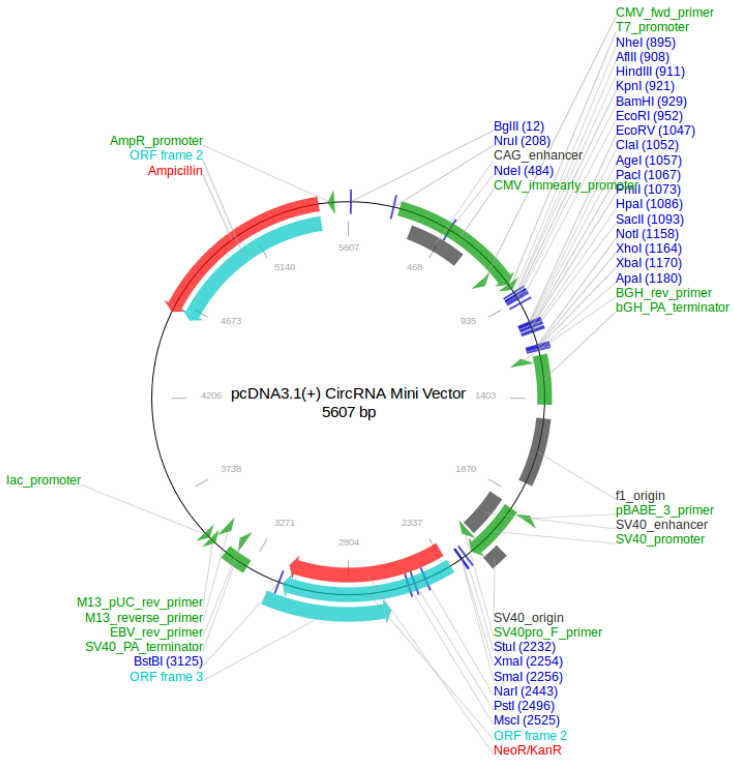

Figure 7.4: pcDNA3.1(+) circRNA miniVector Map. Exons were cloned into this vector to circularize the desired exons. The enzyme sites of ClaI and SacII were used in this experiment.

7.3 Results and discussion

CircRNA landscape in the ageing killifish

The analysis of the pilot data shows that the killifish brain exhibits a greater variety of circRNAs than muscle. This is in agreement with previous findings that the circRNA diversity is greater in the brain of flies and mice compared to other tissues [34, 48]. Genes expressing circRNAs in the brain were enriched for functions of the nervous system, while genes expressing circRNAs in the muscle were enriched for muscle functions (Fig. 7.6). A core set of 422 circRNAs was expressed at all ages (see Figure 7.5b). Additional 203 circRNAs were expressed in both adult and old fish. Only few circRNAs were age-specific. A comparison with human and mouse circRNAs revealed that approximately 13 % of the killifish genes expressing circRNAs also expressed circRNAs in human

and mice (see Figure 7.5c). These 93 genes are of particular interest as they might reveal a common mechanism of circRNAs shared by vertebrates.

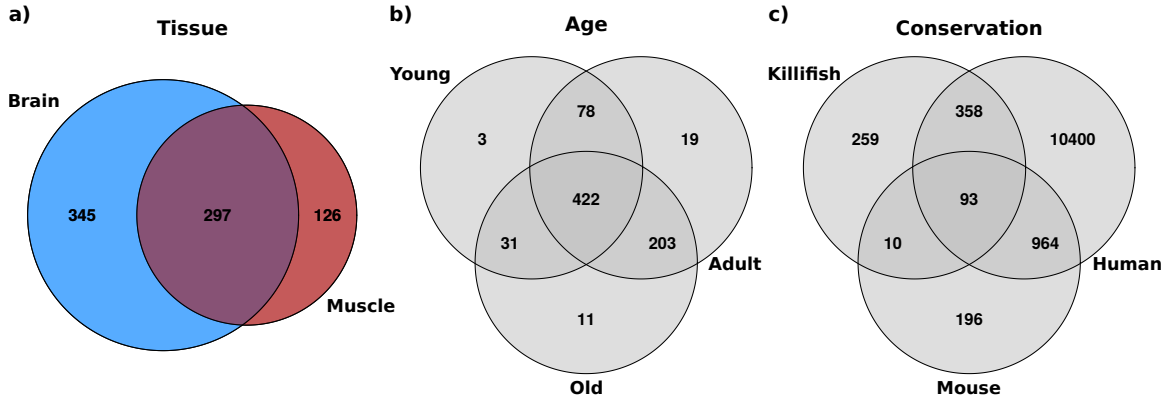

Figure 7.5: Results of the pilot study. a-b) Overlap of circRNAs expressed in different tissues or at different age time points. **c)** Conservation across species: Same host genes across different species that are able to express circular RNAs

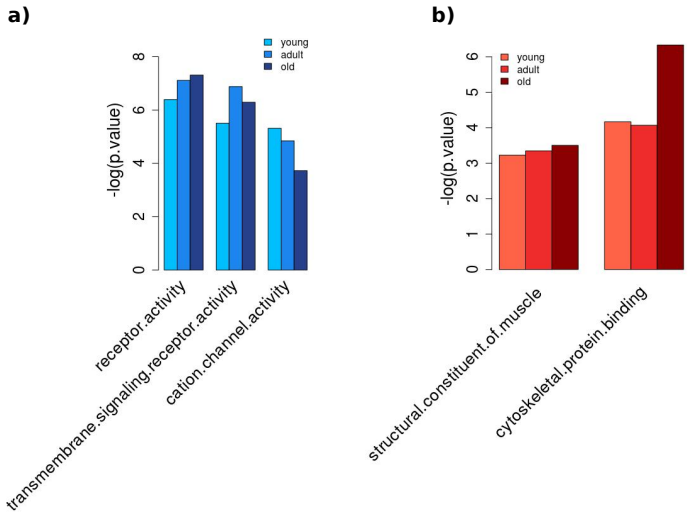

Figure 7.6: Go enrichment of brain and muscle specific circRNAs. GO terms enriched in host genes expressing circRNA in either **a)** brain or **b)** muscle.

Selected circRNAs increase expression with age

As shown in Figure 7.5a only a minor proportion of circRNAs was truly age specific. Hence, it was of great interest whether the expression of the common circRNAs remained stable throughout life or if there are circRNAs whose expression changed during ageing. Therefore, I normalized the circRNA expression with their host gene's expression and clustered the circRNAs based on their relative expression patterns. The k-means clustering showed clusters in which the relative circRNA abundance increased most from young to adult fish, indicating involvement in development (compare Fig. 7.7 Cluster 1), while 253 circRNAs increased most from adult to old fish, suggesting a possible role during ageing (compare Fig. 7.7 Cluster 2 and 8).

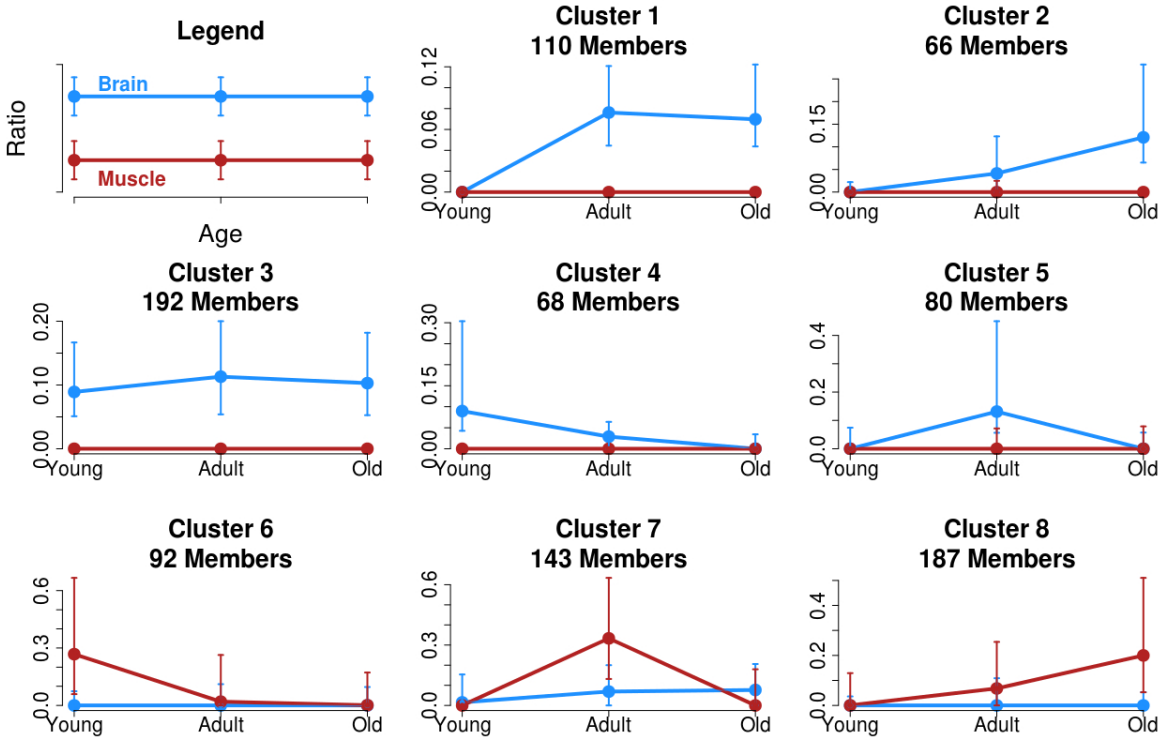

Figure 7.7: Temporal Clustering. Cluster median of the relative circRNA expression for young, adult and old samples. Blue represents the median over all brain circRNAs. Red represents the median over all muscle circRNAs. The errorbars represent the 25th and 75th quantile.

Validation

Thirty-seven circRNAs were selected for validation using qPCR based on the circRNAs' expression patterns (increasing with age), and conservation. qPCR primers were designed to obtain PCR products representing the circRNA junction. Seventeen of these circRNAs were confirmed with high confidence, while only a few could be dismissed due to lack of PCR products. The remaining circRNAs showed a smeared melting curve and it could not be determined if the back-splice junction existed or not. As mentioned in chapter 3.1, a smeared melting cure might indicate overlapping circRNAs. Thus, these circRNAs can neither be confirmed nor disregarded as false positives. Treating the samples with *RNaseR* would be more reliable to validate circRNAs. If the primers also yielded PCR products in the *RNAseR* treated samples the circRNA would be confirmed.

Over-expression strategy

Four (IGF1R, ROBO2, PLCB1, and XPO7) of these 17 highly confident circRNAs were selected for over-expression. They were chosen based on their conservation, their temporal expression pattern in RNA-Seq, their validated temporal expression pattern using qPCR (Fig. 7.8), and biological relevance. The primers were designed for cDNA libraries to only capture the exons for circularization (Fig. 7.9). After using Sanger sequencing to validate the sequence, the PCR products were cloned into the pcDNA3.1(+) circRNA miniVector (Fig. 7.4). The plasmids containing the inserts were cloned into the killifish Tol2 vector, and injected into the killifish embryos during the one-cell stage. Seven days after injection, RNA was extracted from the embryos. Currently, the libraries are being prepared for validation with qPCR. If the pcDNA3.1(+) circRNA miniVector is sufficient to form the desired circRNAs, it would be an indirect proof that the mechanism for exon circularization is conserved across species.

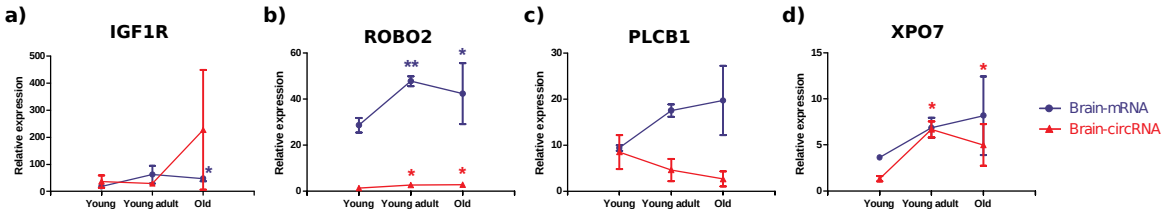

Figure 7.8: qPCR results. qPCR results of the circRNAs that were selected to be cloned into the pcDNA3.1(+) circRNA miniVector. Only data for brain is shown.

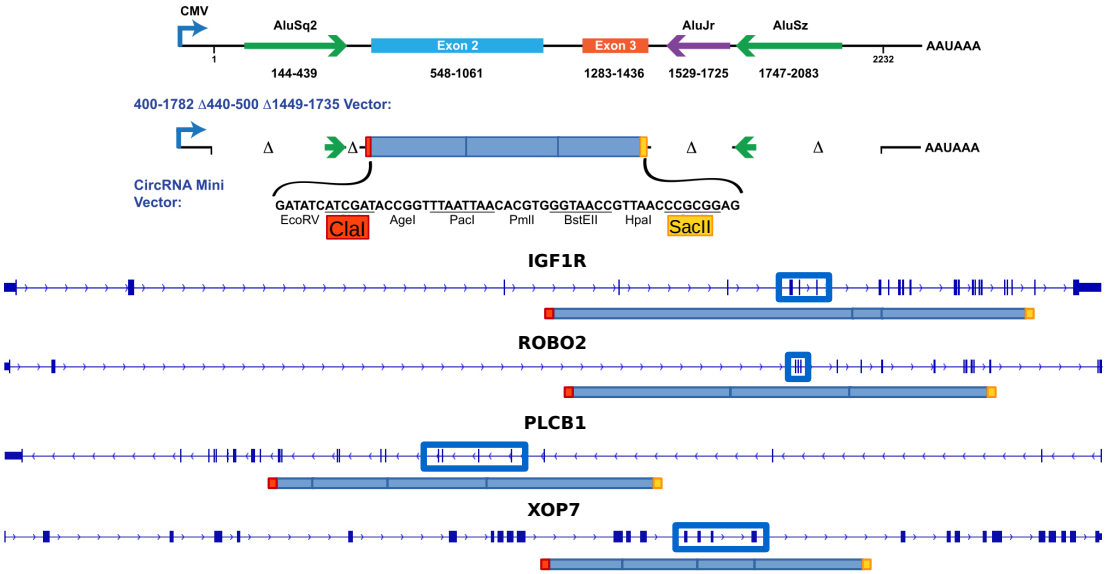

Figure 7.9: Cloning strategy and target regions for circRNA over expression IGF1R, ROBO2, PLCB1, and XPO7 circRNAs were selected to be cloned into the pcDNA3.1(+) circRNA miniVector. The upper picture shows a schematic view of the experimental design by Liang *et al.* [16]. The second line shows a schematic view of the circRNA miniVector cloning site. The enzyme sites for ClaI and SacII were used to cut and paste the exons of IGF1R, ROBO2, PLCB1, and XPO7. The whole gene models of these genes are listed underneath. The blue boxes highlight the circRNA exons chosen for cloning.

7.4 Conclusions

With this pilot study I showed that the killifish is a good model organism to study circRNAs during ageing. Running FUCHS and FUCHS*denovo* on a new dataset of long reads (Table 7.2) will give the opportunity study the circRNA landscape in the ageing killifish in much more detail. The study would benefit from running FUCHS*denovo* as killifish is not as well annotated as the mouse or human genome. Thus performing sequence based computational analysis would only be meaningful after reconstructing the exon-intron model from long reads.

Table 7.2: Sample sheet 2

Sample name	Tissue	Age [w]	Strain	Replicates	Library
Y_muscle_1-2	muscle	3.5	killifish (GRZ-AD)	2	250 BP PE ribo-
YA_muscle_1-2	muscle	8.5	killifish (GRZ-AD)	2	250 BP PE ribo-
O_muscle_1-2	muscle	14.0	killifish (GRZ-AD)	2	250 BP PE ribo-
Y_ brain_ 1-3	brain	3.5	killifish (GRZ-AD)	3	250 BP PE ribo-
YA_ brain_ 1-3	brain	8.5	killifish (GRZ-AD)	3	250 BP PE ribo-
O_ brain_ 1-3	brain	14.0	killifish (GRZ-AD)	3	250 BP PE ribo-

References

[1] Jung C. Lee and Robin R. Gutell. Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. *Journal of molecular biology*, 344(5):1225–1249, December 2004.

[2] Petr Daneck, Christoffer Nellaker, Rebecca E. McIntyre, Jorge E. Buendia-Buendia, Suzannah Bumpstead, Chris P. Ponting, Jonathan Flint, Richard Durbin, Thomas M. Keane, and David J. Adams. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biology*, 13(4):R26+, 2012.

[3] Sheng Li and Christopher E. Mason. The Pivotal Regulatory Landscape of RNA Modifications. *Annual Review of Genomics and Human Genetics*, 15(1):127–150, 2014.

[4] Neema Agrawal, P. V. N. Dasaradhi, Asif Mohmmed, Pawan Malhotra, Raj K. Bhatnagar, and Sunil K. Mukherjee. RNA Interference: Biology, Mechanism, and Applications. *Microbiology and Molecular Biology Reviews*, 67(4):657–685, December 2003.

[5] Maden and JohnM Hughes. Eukaryotic ribosomal RNA: the recent excitement in the nucleotide modification problem. 105(7-8):391–400, 1997.

[6] Yehoshua Enuka, Mattia Lauriola, Morris E. Feldman, Aldema Sas-Chen, Igor Ulitsky, and Yosef Yarden. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. *Nucleic Acids Research*, 44(3):1370–1383, December 2015.

[7] Peter L. Wang, Yun Bao, Muh-Ching Yee, Steven P. Barrett, Gregory J. Hogan, Mari N. Olsen, José R. Dinneny, Patrick O. Brown, and Julia Salzman. Circular RNA Is Expressed across the Eukaryotic Tree of Life. *PLOS ONE*, 9(3):e90859+, March 2014.

[8] Julia Salzman, Charles Gawad, Peter L. Wang, Norman Lacayo, and Patrick O. Brown. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. *PLOS ONE*, 7(2):e30733+, February 2012.

[9] ming-Ta Hsu and Miguel Coca-Prados. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature*, 280(5720):339–340, July 1979.

[10] Janice M. Nigro, Kathleen R. Cho, Eric R. Fearon, Scott E. Kern, Ruppert, Jonathan D. Oliner, Kenneth W. Kinzler, and Bert Vogelstein. Scrambled exons. *Cell*, 64(3):607–613, February 1991.

[11] C. Cocquerelle, P. Daubersies, M. A. Majérus, J. P. Kerckaert, and B. Bailleul. Splicing with inverted order of exons occurs proximal to large introns. *The EMBO journal*, 11(3):1095–1098, March 1992.

[12] B. Bailleul. During in vivo maturation of eukaryotic nuclear mRNA, splicing yields excised exon circles. *Nucleic Acids Research*, 24(6):1015–1019, March 1996.

[13] Blanche Capel, Amanda Swain, Silvia Nicolis, Adam Hacker, Michael Walter, Peter Koopman, Peter Goodfellow, and Robin Lovell-Badge. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell*, 73(5):1019–1030, June 1993.

[14] Reut Ashwal-Fluss, Markus Meyer, Nagarjuna R. Pamudurti, Andranik Ivanov, Osnat Bartok, Mor Hanan, Naveh Evantal, Sebastian Memczak, Nikolaus Rajewsky, and Sebastian Kadener. circRNA Biogenesis Competes with Pre-mRNA Splicing. *Molecular Cell*, 56(1):55–66, October 2014.

[15] Stefan Starke, Isabelle Jost, Oliver Rossbach, Tim Schneider, Silke Schreiner, Lee-Hsueh Hung, and Albrecht Bindereif. Exon Circularization Requires Canonical Splice Signals. *Cell Reports*, 10(1):103–111, January 2015.

[16] Dongming Liang and Jeremy E. Wilusz. Short intronic repeat sequences facilitate circular RNA production. *Genes & Development*, 28(20):2233–2247, October 2014.

[17] J. M. Houseley, Z. Garcia-Casado, M. Pascual, N. Paricio, K. M. C. O'Dell, D. G. Monckton, and R. D. Artero. Noncanonical RNAs From Transcripts of the Drosophila muscleblind Gene. *Journal of Heredity*, 97(3):253–260, May 2006.

[18] Simon J. Conn, Katherine A. Pillman, John Toubia, Vanessa M. Conn, Marika Salmanidis, Caroline A. Phillips, Suraya Roslan, Andreas W. Schreiber, Philip A. Gregory, and Gregory J. Goodall. The RNA Binding Protein Quaking Regulates Formation of circRNAs. *Cell*, 160(6):1125–1134, March 2015.

[19] Andranik Ivanov, Sebastian Memczak, Emanuel Wyler, Francesca Torti, Hagit T. Porath, Marta R. Orejuela, Michael Piechotta, Erez Y. Levanon, Markus Landthaler, Christoph Dieterich, and Nikolaus Rajewsky. Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals. *Cell Reports*, 10(2):170–177, January 2015.

[20] Yang Zhang, Wei Xue, Xiang Li, Jun Zhang, Siye Chen, Jia-Lin Zhang, Li Yang, and Ling-Ling Chen. The Biogenesis of Nascent Circular RNAs. *Cell Reports*, 15(3):611–624, April 2016.

[21] Thomas B. Hansen, Erik D. Wiklund, Jesper B. Bramsen, Sune B. Villadsen, Aaron L. Statham, Susan J. Clark, and Jørgen Kjems. miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *The EMBO Journal*, 30(21):4414–4422, November 2011.

[22] Thomas B. Hansen, Trine I. Jensen, Bettina H. Clausen, Jesper B. Bramsen, Bente Finsen, Christian K. Damgaard, and Jørgen Kjems. Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441):384–388, February 2013.

[23] Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D. Mackowiak, Lea H. Gregersen, Mathias Munschauer, Alexander Loewer, Ulrike Ziebold, Markus Landthaler, Christine Kocks, Ferdinand le Noble, and Nikolaus Rajewsky. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441):333–338, February 2013.

[24] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141, April 2010.

[25] Alexander G. Baltz, Mathias Munschauer, Björn Schwanhäusser, Alexandra Vasile, Yasuhiro Murakawa, Markus Schueler, Noah Youngs, Duncan Penfold-Brown, Kevin Drew, Miha Milek, Emanuel Wyler, Richard Bonneau, Matthias Selbach, Christoph Dieterich, and Markus Landthaler. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*, 46(5):674–690, June 2012.

[26] Xianwen Meng, Xue Li, Peijing Zhang, Jingjing Wang, Yincong Zhou, and Ming Chen. Circular RNA: an emerging key player in RNA world. *Briefings in bioinformatics*, June 2016.

[27] Xuning Wang, Yue Zhang, Liang Huang, Jiajin Zhang, Fei Pan, Bing Li, Yongfeng Yan, Baoqing Jia, Hongyi Liu, Shiyou Li, and Wei Zheng. Decreased expression of hsa_circ_001988 in colorectal cancer and its clinical significances. *International journal of clinical and experimental pathology*, 8(12):16020–16025, 2015.

[28] Peifei Li, Shengcan Chen, Huilin Chen, Xiaoyan Mo, Tianwen Li, Yongfu Shao, Bingxiu Xiao, and Junming Guo. Using circular RNA as a novel type of biomarker in the screening of gastric cancer. *Clinica Chimica Acta*, 444:132–136, April 2015.

[29] Christin E. Burd, William R. Jeck, Yan Liu, Hanna K. Sanoff, Zefeng Wang, and Norman E. Sharpless. Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk. *PLOS Genetics*, 6(12):e1001233+, December 2010.

[30] Walter J. Lukiw. Circular RNA (circRNA) in Alzheimer's disease (AD). *Frontiers in Genetics*, 4, 2013.

[31] Jun Cheng, Franziska Metge, and Christoph Dieterich. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics*, 32(7):1094–1096, April 2016.

[32] Franziska Metge, Lisa F. Czaja-Hasse, Richard Reinhardt, and Christoph Dieterich. FUCHS-towards full circular RNA characterization using RNAseq. *PeerJ*, 5, 2017.

[33] William R. Jeck and Norman E. Sharpless. Detecting and characterizing circular RNAs. *Nat Biotechnol*, 32(5):453–461, May 2014.

[34] Jakub O. Westholm, Pedro Miura, Sara Olson, Sol Shenker, Brian Joseph, Piero Sanfilippo, Susan E. Celniker, Brenton R. Graveley, and Eric C. Lai. Genome-wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation. *Cell Reports*, 9(5):1966–1980, December 2014.

[35] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, March 2012.

[36] Steve Hoffmann, Christian Otto, Gero Doose, Andrea Tanzer, David Langenberger, Sabina Christ, Manfred Kunz, Lesca M. Holdt, Daniel Teupser, Jörg Hackermüller, and Peter F. Stadler. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biology*, 15(2):R34+, 2014.

[37] Xiao-Ou Zhang, Hai-Bin Wang, Yang Zhang, Xuhua Lu, Ling-Ling Chen, and Li Yang. Complementary Sequence-Mediated Exon Circularization. *Cell*, 159(1):134–147, September 2014.

[38] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[39] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, March 2010.

[40] Yuan Gao, Jinfeng Wang, and Fangqing Zhao. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biology*, 16(1):4+, 2015.

[41] Linda Szabo, Robert Morey, Nathan J. Palpant, Peter L. Wang, Nastaran Afari, Chuan Jiang, Mana M. Parast, Charles E. Murry, Louise C. Laurent, and Julia Salzman. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biology*, 16(1):126+, June 2015.

[42] Matthias Dodt, Johannes Roehr, Rina Ahmed, and Christoph Dieterich. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, 1(3):895–905, December 2012.

[43] Morris A. Jette, Andy B. Yoo, and Mark Grondona. SLURM: Simple Linux Utility for Resource Management. In *In Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP) 2003*, pages 44–60, 2002.

[44] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48+, February 2009.

[45] Xiao-Ou Zhang, Rui Dong, Yang Zhang, Jia-Lin Zhang, Zheng Luo, Jun Zhang, Ling-Ling Chen, and Li Yang. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Research*, 26(9):1277–1287, September 2016.

[46] Yuan Gao, Jinfeng Wang, Yi Zheng, Jinyang Zhang, Shuai Chen, and Fangqing Zhao. Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nature Communications*, 7:12060+, June 2016.

[47] Naoko Abe, Ken Matsumoto, Mizuki Nishihara, Yukiko Nakano, Aya Shibata, Hideto Maruyama, Satoshi Shuto, Akira Matsuda, Minoru Yoshida, Yoshihiro Ito, and Hiroshi Abe. Rolling Circle Translation of Circular RNA in Living Human Cells. *Scientific Reports*, 5:16435+, November 2015.

[48] Xintian You, Irena Vlatkovic, Ana Babic, Tristan Will, Irina Epstein, Georgi Tushev, Güney Akbalik, Mantian Wang, Caspar Glock, Claudia Quedenau, Xi Wang, Jingyi Hou, Hongyu Liu, Wei Sun, Sivakumar Sambandan, Tao Chen, Erin M. Schuman, and Wei Chen. Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nature Neuroscience*, 18(4):603–610, February 2015.

[49] Charles E. Grant, Timothy L. Bailey, and William S. Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011.

[50] Junjie U. Guo, Vikram Agarwal, Huili Guo, and David P. Bartel. Expanded identification and characterization of mammalian circular RNAs. *Genome Biology*, 15(7):409+, July 2014.

[51] T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, May 2011.

[52] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24+, February 2007.

[53] Vittorio Calderone, Javier Gallego, Gonzalo Fernandez-Miranda, Ester Garcia-Pras, Carlos Maillo, Annalisa Berzigotti, Marc Mejias, Felice-Alessio Bava, Ana Angulo-Urarte, Mariona Graupera, Pilar Navarro, Jaime Bosch, Mercedes Fernandez, and Raul Mendez. Sequential Functions of CPEB1 and CPEB4 Regulate Pathologic Expression of Vascular Endothelial Growth Factor and Angiogenesis in Chronic Liver Disease. *Gastroenterology*, 150(4):982–997.e30, April 2016.

[54] Sean B. Carroll. Homeotic genes and the evolution of arthropods and chordates. *Nature*, 376(6540):479–485, August 1995.

[55] Joseph C. Pearson, Derek Lemons, and William McGinnis. Modulating Hox gene functions during animal body patterning. *Nature Reviews Genetics*, 6(12):893–904, November 2005.

[56] Theresia R. Kress, Ian G. Cannell, Arjan B. Brenkman, Birgit Samans, Matthias Gaestel, Paul Roepman, Boudewijn M. Burgering, Martin Bushell, Andreas Rosenwald, and Martin Eilers. The MK5/PRAK Kinase and Myc Form a Negative Feedback Loop that Is Disrupted during Colorectal Tumorigenesis. *Molecular Cell*, 41(4):445–457, February 2011.

[57] Ordan J. Lehmann, Jane C. Sowden, Peter Carlsson, Tim Jordan, and Shomi S. Bhattacharya. Fox's in development and disease. *Trends in genetics : TIG*, 19(6):339–344, June 2003.

[58] Kun Wang and Pei-Feng Li. Foxo3a Regulates Apoptosis by Negatively Targeting miR-21. *Journal of Biological Chemistry*, 285(22):16958–16966, May 2010.

[59] Brian J. Morris, Donald Craig C. Willcox, Timothy A. Donlon, and Bradley J. Willcox. FOXO3: A Major Gene for Human Longevity–A Mini-Review. *Gerontology*, 61(6):515–525, 2015.

[60] William W. Du, Weining Yang, Yu Chen, Zhong-Kai Wu, Francis S. Foster, Zhenguo Yang, Xiangmin Li, and Burton B. Yang. Foxo3 circular RNA promotes cardiac senescence by modulating multiple factors associated with stress and senescence responses. *European Heart Journal*, pages ehw001+, February 2016.

[61] Agnieszka Rybak-Wolf, Christin Stottmeister, Petar Glažar, Marvin Jens, Natalia Pino, Sebastian Giusti, Mor Hanan, Mikaela Behm, Osnat Bartok, Reut Ashwal-Fluss, Margareta Herzog, Luisa Schreyer, Panagiotis Papavasileiou, Andranik Ivanov, Marie Öhman, Damian Refojo, Sebastian Kadener, and Nikolaus Rajewsky. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular Cell*, 58(5):870–885, June 2015.

[62] Morten T. Venø, Thomas B. Hansen, Susanne T. Venø, Bettina H. Clausen, Manuela Grebing, Bente Finsen, Ida E. Holm, and Jørgen Kjems. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. *Genome Biology*, 16(1):1–17, November 2015.

[63] Eva Terzibasi, Dario R. Valenzano, and Alessandro Cellerino. The short-lived fish Nothobranchius furzeri as a new model system for aging studies. *Experimental Gerontology*, 42(1-2):81–89, January 2007.

[64] Yumi Kim, Hong G. Nam, and Dario R. Valenzano. The short-lived African turquoise killifish: an emerging experimental model for ageing. *Disease Models & Mechanisms*, 9(2):115–129, February 2016.

[65] Itamar Harel, Bérénice A. Benayoun, Ben Machado, Param P. Singh, Chi-Kuo Hu, Matthew F. Pech, Dario R. Valenzano, Elisa Zhang, Sabrina C. Sharp, Steven E. Artandi, and Anne Brunet. A Platform for Rapid Exploration of Aging and Diseases in a Naturally Short-Lived Vertebrate. *Cell*, 160(5):1013–1026, February 2015.

[66] Petar Glažar, Panagiotis Papavasileiou, and Nikolaus Rajewsky. circBase: a database for circular RNAs. *RNA*, September 2014.

[67] Dario R. Valenzano, Sabrina Sharp, and Anne Brunet. Transposon-Mediated Transgenesis in the Short-Lived African Killifish Nothobranchius furzeri, a Vertebrate Model for Aging. *G3: Genes, Genomes, Genetics*, 1(7):531–538, December 2011.

Erklärung zur Dissertation

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt habe, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universtität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.
Die Bestimungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Dr. Dario Valenzano betreut worden.
Nachfolgend genannte Teilpublikationen liegen vor:

1. Cheng, J., Metge, F., and Dieterich, C. (2016). Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics*, 32(7):1094–1096. [31]

2. Metge, F., Czaja-Hasse, L. F., Reinhardt, R., and Dieterich, C. (2017). FUCHS-towards full circular RNA characterization using RNAseq. *PeerJ*, 5. [32]

Franziska Metge
31. July 2018