Development of a circular RNA detection pipeline and its application in medulloblastoma

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine Universität Düsseldorf

vorgelegt von
Daniel Rickert
aus Borken, NRW

Düsseldorf, August 2022

Aus der Klinik für Kinder-Onkologie, -Hämatologie und Klinische Immunologie sowie dem Institut für Neuropathologie der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:
1. Prof. Dr. Gunnar Klau
2. Prof. Dr. Guido Reifenberger

Tag der mündlichen Prüfung: 15.06.2023

Chapters

Summary.......6

Zusammenfassung.......8

1 Introduction.......10

- 1.1 Pediatric malignancies.......10
- 1.2 Medulloblastoma.......11
  - 1.2.0 Overview.......11
  - 1.2.1 History.......12
  - 1.2.2 Clinical features.......12
  - 1.2.3 Epidemiology.......13
  - 1.2.4 WNT medulloblastoma.......14
  - 1.2.5 SHH medulloblastoma.......14
  - 1.2.6 Group 3 medulloblastoma.......15
  - 1.2.7 Group 4 medulloblastoma.......16
  - 1.2.8 Medulloblastoma risk stratification.......17
- 1.3 RNA sequencing.......17
- 1.4 Circular RNA.......19
  - 1.4.0 Overview.......19
  - 1.4.1 Classification of circular RNAs.......20
  - 1.4.2 Synthesis of circular RNAs.......20
  - 1.4.3 Functions of circular RNAs.......21
  - 1.4.4 Circular RNAs as disease-specific biomarker.......22
- 1.5 Detection of circular RNAs.......23
  - 1.5.0 Overview.......23
  - *1.5.1* Detection of circular RNAs *in silico* based on RNA sequencing data.......24
  - *1.5.2* Detection of circular RNAs *in vitro / in vivo*.......25
- 1.6 Circular RNA and medulloblastoma.......25
- 1.7 Other tools for circular RNA research.......26
- 1.8 Snakemake.......26

2 Aim of this thesis....28

3 Materials and methods....30

3.1 Wet lab experiments....30

3.1.1 Patient samples....30

3.1.2 *MYC*/RNAseR cell line models....30

3.1.3 Lentiviral vector construction for *MYC* overexpression....31

3.1.4 *MYC*/RNAseR RNA isolation and sample preparation....31

3.2 Dry lab experiments....31

3.2.1 Software packages....32

3.2.2 Non-default Perl packages....32

3.2.3 Non-default R packages....32

3.2.4 Hardware used....33

3.2.5 RNA sequencing pre-processing....33

3.2.6 Circular RNA detection using the *circs* workflow....33

3.2.7 Data processing....34

3.2.8 MiOncoCircDB overlap and comparisons....34

4 Results....36

4.1 Development and application of the *circs* workflow....36

4.1.1 Selection of appropriate circular RNA detection pipelines....36

4.1.2 *Circs* internal data flow....38

4.1.3 Execution of *circs*....40

4.1.4 Circs_snake....42

4.1.5 Circs_snake execution....47

4.2 Discovery cohort....48

4.2.1 Detection of circular RNAs in the discovery cohort....48

4.2.2 Medulloblastoma group classification utilizing circular RNA data in the discovery cohort....51

4.3 Validation cohort....54

4.3.1 Medulloblastoma group classification utilizing circular RNA data in the validation

cohort......54
4.4 Cross-dataset circular RNA signal evaluation......56
4.5 Medulloblastoma pan-cohort group circular RNA biomarker definition......59
4.5.1 Circ*RMST* as WNT medulloblastoma biomarker......62
4.5.2 Circ*ISPD* as SHH medulloblastoma biomarker......65
4.5.3 Circ*EXOC6B* as Group 4 medulloblastoma biomarker......66
4.5.4 Other medulloblastoma group-specific circular RNA biomarkers......68
4.6 Medulloblastoma-subtype-specific circular RNA biomarkers......69
4.6 Validation of *in silico* prediction based on circs workflow......72
4.6.1 Database-aided validation of circRNAs detected with *circs*......72
4.6.2 Quantitative circular RNA *in silico* specificity determination......74
4.6.3 Circleseq based circular RNA validation in an *in vitro* model......77
4.6.4 *In silico* validation of experimental data......79
4.6.5 RNAseR validation rate assessment......81
4.6.6 Cell-line-specific RNAseR validation rate......84
4.6.7 Circular RNA pertubations in the *MYC*/RNAseR dataset......86
5 Discussion......88
6 References......96
7 Appendix......105
7.1 Supplementary figures......105
7.2 Supplementary tables......109
7.3 Figure directory......114
7.4 List of tables......120
Curriculum Vitae......121
Publications and presentations......123
Acknowledgments......124
Affirmation......125

Summary

Recent advances in next generation sequencing have provided a rich resource of large-scale RNA sequencing (RNA-Seq) data from various types of human cancer, thereby fostering omics-based cancer research and biomarker discovery. However, the instability of linear mRNA limits its use for robust biomarker detection in routine diagnostic applications. This thesis therefore evaluates the potential of a more stable RNA species, namely circular RNA (circRNA), which shows specific expression patterns according to developmental stage and differentiation of cells and tissues. Circular RNA is a closed loop of single-stranded RNA and amounts to ~1% of total RNA detectable in a given sample. Currently, there are several circRNA quantification methods publicly available. To accelerate circRNA research and overcome certain single-pipeline-based limitations, a novel multi-pipeline circRNA detection method called '*circs*' was developed. This method allows highly sensitive circRNA detection and achieves a lower false-positive rate compared to previous circRNA-detection pipelines. In this work, 'circ' was successfully applied to both published and unpublished RNA-Seq data and allowed to quantify circRNA expression in RNAseq data sets from two independent cohorts of medulloblastoma (MB), the most common type of malignant brain tumor in children. These cohorts consisted of RNAseq data sets of 38 and 35 MB patient samples, and included various proportions of the four major MB groups: wingless (WNT), sonic hedgehog (SHH), group 3 and group 4. They are characterized by group-specific recurrent genomic alterations leading to aberrant activation of distinct signaling pathways and divergent clinical behavior concerning likelihood of metastasis formation and patient outcome.

The work summarized in this thesis shows that circRNA expression in MB tissue samples can be used to precisely classify tumors into these distinct MB groups without any additional data. In fact, MB group assignment based on circRNA expression profiles proved as precise as assignment based on Similarity Network Fusion (SNF), which integrates multiple omics layers for molecular tumor classification. CircRNA expression profiles differed significantly between the MB groups and were validated in the two

independent patient cohorts. The validated MB-group-specific circRNA profiles not only allowed reliable distinction between groups, but also identified individual circRNA species with selective expression in single MB groups. For example, circ*RMST* was found to be a remarkably stable and highly expressed biomarker for WNT MB in both cohorts, circ*ISPD* was the top biomarker for SHH MB, and circ*EXOC6B* was specifically upregulated in Group 4 MB. Employing additional online tumor databases, it was possible to confirm these circRNA biomarkers in further published datasets. Additionally, the specificity of circ*RMST* upregulation for WNT MB was substantiated using circRNA expression profiles of >2000 tissue samples, including various other cancer entities and control tissues.

To further validate several circRNA biomarker candidates, the ‘circleseq’ protocol was used with isogenic *MYC*-overexpressing MB cell lines. In addition to identifying many circRNAs detected in human MB tissue samples, these results also confirmed a previously observed trend of *MYC* overexpression being an indicator of globally decreased circRNA abundance in MB. The ‘circs’ pipeline developed in this thesis is freely available for public use, thus enabling other researchers to re-analyze their RNA-Seq data to uncover another omics data layer with highly promising biomarker potential.

Zusammenfassung

Die großen Fortschritte in der Next-Generation-Sequenzierung führten zu umfangreichen RNA-Sequenzierungsdaten (RNA-Seq) von unterschiedlichen Krebsarten beim Menschen und unterstützen damit die Omics-basierte Krebsforschung und die Entdeckung von neuen molekularen Biomarkern. Die Instabilität der linearen mRNA schränkt jedoch die Verwendung von mRNA-basierten Signaturen für den robusten Biomarker-Nachweis im diagnostischen Alltag ein. Die vorliegende Dissertationsarbeit beschäftigt sich daher mit dem diesbezüglichen Potenzial einer stabileren RNA-Spezies, nämlich der zirkulären RNA (circRNA), die je nach Entwicklungsstadium und Differenzierung von Zellen und Geweben sehr spezifische Expressionsmuster zeigt. Zirkuläre RNA stellt eine geschlossene Schleife aus einzelsträngiger RNA dar und macht insgesamt nur etwa 1 % der in einem Zell- oder Gewebe-basierten Extrakt nachweisbaren Gesamt-RNA aus. Aktuell sind mehrere circRNA-Quantifizierungsmethoden öffentlich verfügbar. Um die circRNA-basierte Forschung weiter zu beschleunigen und bestehende Limitationen bei Verwendung einer einzelnen der verfügbaren Analysepipelines zu überwinden, wurde in dieser Dissertationsarbeit eine neuartige Multi-Pipeline-Detektionsmethode für circRNA namens „circ“ entwickelt. Diese Pipeline ist hochempfindlich für den circRNA-Nachweis und erreicht eine geringere Falsch-Positiv-Rate im Vergleich zu den bislang vorhandenen circRNA-Nachweispipelines. In dieser Arbeit wurde 'circ' erfolgreich auf veröffentlichte sowie unveröffentlichte RNA-Seq-Daten angewendet und ermöglichte die Quantifizierung der circRNA-Expression in RNAseq-Datensätzen aus zwei unabhängigen Kohorten von Medulloblastomen (MB), der häufigsten Art von bösartigen Hirntumoren im Kindesalter. Diese Kohorten bestanden aus RNAseq-Datensätzen von 38 und 35 MB Patientenproben, die jeweils Proben der vier großen MB-Gruppen in unterschiedlichen Anteilen enthielten. Diese MB-Gruppen werden Wingless (WNT), Sonic Hedgehog (SHH), Gruppe 3 und Gruppe 4 genannt. Sie zeichnen sich durch gruppenspezifische genomische Veränderungen aus, die zu einer aberranten Aktivierung unterschiedlicher Signalwege und einem unterschiedlichen klinischen Verhalten hinsichtlich der Wahrscheinlichkeit einer Metastasenbildung und der Prognose für die Patienten führen.

Die in dieser Dissertationsschrift zusammengefassten Arbeiten zeigen, dass die circRNA-Expression in MB-Gewebeproben verwendet werden kann, um die Tumoren sehr präzise in die einzelnen MB-Gruppen einzuordnen. Tatsächlich war die Zuordnung der MB-Gruppen basierend auf circRNA-Expressionsprofilen genauso präzise wie die Zuordnung basierend auf der Similarity Network Fusion (SNF) Analyse, die mehrere Omics-Datensätze für die molekulare Tumorklassifizierung integriert. Insgesamt unterschieden sich die circRNA-Expressionsprofile signifikant zwischen den MB-Gruppen, was in den beiden unabhängigen Patientenkohorten validiert wurde. Die validierten circRNA-Profile ermöglichten nicht nur eine zuverlässige Unterscheidung der Gruppen sondern identifizierten auch einzelne circRNA-Spezies mit selektiver Expression in den einzelnen MB-Gruppen. Dabei erwies sich circ*RMST* in beiden Kohorten als stabil und stark exprimierter Biomarker für Medulloblastome der WNT-Gruppe. Im Gegensatz dazu war circI*SPD* der wichtigste Biomarker für Medulloblastome der SHH-Gruppe, während circ*EXOC6B* in den Gruppe 4 Medulloblastomen spezifisch hochreguliert war. Durch die Analyse zusätzlicher Tumordatenbanken konnten diese circRNA-Biomarker in weiteren öffentlich verfügbaren Datensätzen bestätigt werden. Darüber hinaus wurde die Spezifität der circ*RMST*-Hochregulation in WNT-Medulloblastomen anhand von circRNA-Expressionsprofilen in >2000 Gewebeproben einschließlich diverser Krebsentitäten und Kontrollgewebe belegt.

Um weitere circRNA-Biomarkerkandidaten weiter zu validieren, wurde das Protokoll „circleseq“ mit isogenen MYC-überexprimierenden MB-Zelllinien verwendet. Diese Ergebnisse bestätigten nicht nur viele circRNAs, die in menschlichen MB-Gewebeproben mit „circs“ nachgewiesen wurden, sondern validierten auch einen zuvor beobachteten Trend, dass eine MYC-Überpression einen Indikator für eine generell verminderte circRNA-Expression in darstellt. Die hier entwickelte Pipeline „circs“ ist frei verfügbar und öffentlich zugänglich, sodass andere Forscher*innen ihre RNA-Seq-Daten erneut analysieren können, um eine weitere Omics-Datenschicht mit viel versprechendem Biomarkerpotenzial untersuchen zu können.

1 Introduction

1.1 Pediatric malignancies

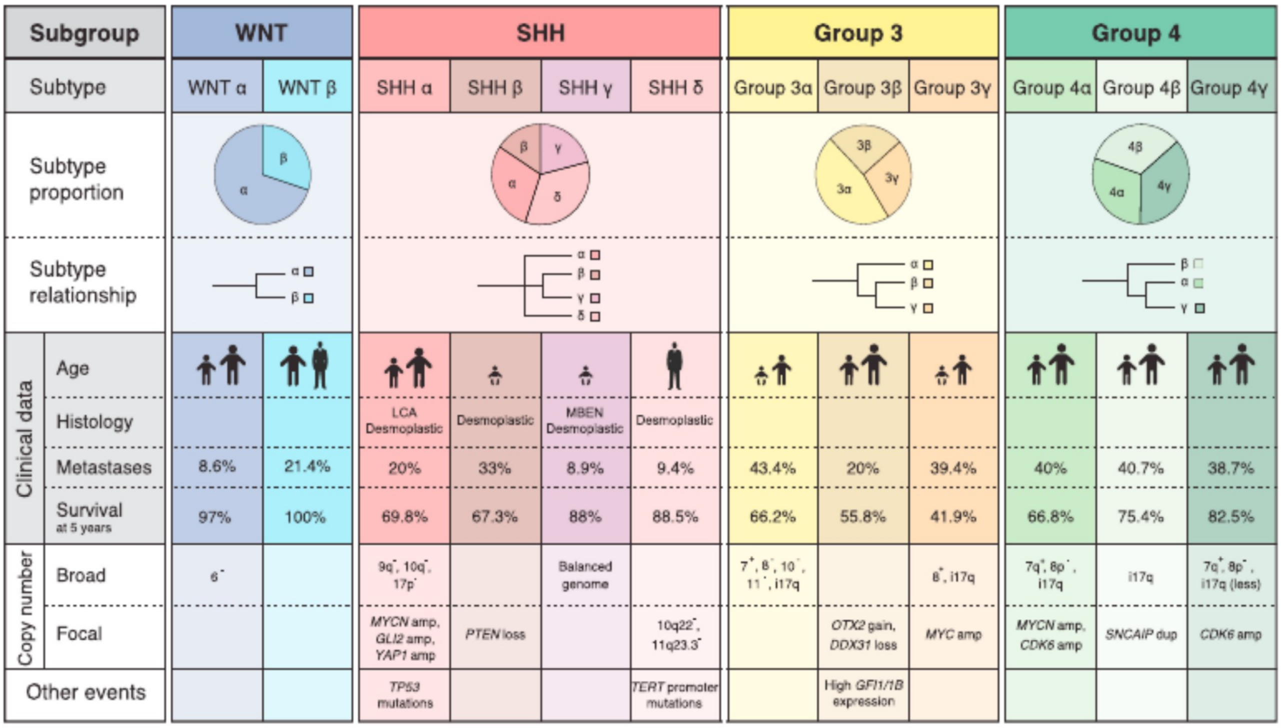
Cancer is a devastating disease that is typically not caused by one single genetic event, but rather arises through an accumulation of alterations in the affected cells. The "hallmarks of cancer" summarize these features[1]: Development of any somatic cell into cancer requires, among other characteristics, abnormal and sustained cell growth that might be caused by mutations in pro-proliferative signaling pathways. This usually is kept under control by growth suppressors that the developing cancer must evade in order to survive. Cell death needs to be avoided by mechanisms that render the cancer cells immortal, making indefinite proliferation possible. The immune system can also destroy cells proliferating in an uncontrolled manner — another housekeeping mechanism to which the growing cancer must adapt. Any living cell needs nutrients, and thus after a certain degree of tumor growth, angiogenesis must be induced to ensure the delivery of blood to the tumor tissue. For solid tumors, a deregulation of the energy metabolism in malignant cells deep inside the tumor can render them more resistant to stressful conditions, such as nutrient deprivation, if angiogenesis is insufficient. Tumors also can spread and form metastases in distant tissues, the last hallmark of cancer[1]. Each of these steps may be achieved due to pre-existing germline mutations and/or acquired somatic mutations, changes in gene expression, epigenetic alterations, or other mechanisms.

Besides leukemias and a number of solid cancers originating outside the central nervous system (CNS), there are several types of malignant brain tumors threatening the lives of children[2–4]. Even state-of-the art medicine does not guarantee survival nor complete recovery in these malignant neoplasms. In general, brain cancers are difficult to treat because the blood-brain barrier complicates drug delivery[5,6], the position of the malignancy often prohibits complete tumor resection due to proximity to vital parts of the CNS and sequelae of treatment are manifold, often resulting in impaired neuro-cognitive abilities in brain cancer survivors[7,8]. If survival of the initial malignancy is achieved, metastasis can eventually threaten the patients life yet again. Immediate or long-term treatment related side-effects from radiotherapy, chemotherapy and/or tumor resection can emerge[2,9–12]. One form of pediatric brain cancer, medulloblastoma (MB), is the most common primary malignant CNS tumor in children, diagnosed ~1000 times per year globally[13(p8)].

1.2 Medulloblastoma

1.2.0 Overview

Medulloblastoma (MB) is the most common malignant brain tumor in children[9]. The sub-categorization of MB into different groups[14,15] in order to improve therapy[16] resulted in four core MB groups: WNT (wingless), SHH (sonic hedgehog), Group 3 and Group 4[17,18] - each with biological distinct and clinically relevant features[19,20]. 5-year survival is the best in WNT MBs, while the prognosis for Group 3 MB is the worst. Group 4 and SHH are both considered intermediate risk. However, this can be assessed more precisely when considering subtypes (see Figure 1). Recommendations have been made to de-escalate therapy for WNT MB patients since the current standard-of-care leads to significant side-effects and suboptimal development of treated children[7,21,20,22]. However, recent studies tried to divide the current four MB groups into more subtypes[18] based on multiple tumor tissue characteristics, leading to a more accurate disease classification and enabling a more targeted approach.



Subgroup		WNT		SHH				Group 3			Group 4		
Subtype		WNT α	WNT β	SHH α	SHH β	SHH γ	SHH δ	Group 3α	Group 3β	Group 3γ	Group 4α	Group 4β	Group 4γ
Subtype proportion													
Subtype relationship													
Clinical data	Age												
	Histology			LCA Desmoplastic	Desmoplastic	MBEN Desmoplastic	Desmoplastic						
	Metastases	8.6%	21.4%	20%	33%	8.9%	9.4%	43.4%	20%	39.4%	40%	40.7%	38.7%
	Survival at 5 years	97%	100%	69.8%	67.3%	88%	88.5%	66.2%	55.8%	41.9%	66.8%	75.4%	82.5%
Copy number	Broad	6^-		$9q^-$, $10q^-$, $17p^-$		Balanced genome		7^+, 8^-, 10^-, 11^-, i17q		8^+, i17q	$7q^+$, $8p^-$, i17q	i17q	$7q^+$, $8p^-$, i17q (less)
	Focal			*MYCN* amp, *GLI2* amp, *YAP1* amp	*PTEN* loss		$10q22^-$, $11q23.3^-$		*OTX2* gain, *DDX31* loss	*MYC* amp	*MYCN* amp, *CDK6* amp	*SNCAIP* dup	*CDK6* amp
Other events				*TP53* mutations			*TERT* promoter mutations		High *GFI1/1B* expression				

Figure 1: Medulloblastoma groups and subtypes displayed with characteristics of each subtype. Taken from Cavalli et.al. 2017.

1.2.1 History

Medulloblastoma was first described in 1926[23] as a malignancy thought to originate from a neuronal progenitor cell type called medulloblasts that - as known today - does not exist[13]. The exact cell type of origin is still a focus of research[24]. In the years that followed, it was reported that radiation improved the patients' survival, as did chemotherapy[23]. Thereafter, recurrent mutations were identified in SHH and WNT pathway genes, paving the way for MB group definitions that reached a consensus of four MB groups in 2012[25]. Subsequently, subtypes were also defined, and continue to be fine-tuned today[3,6,7,21,22,24,25,26(p5)].

1.2.2 Clinical features

Clinical features of MB include unspecific symptoms, such as headache and vomiting, as well as more specific ones, such as ataxia, motor or vision problems that can worsen over time[9]. If a child already has spinal metastasis, the symptoms may also include back pain, gait difficulties and sensory deficits. Medulloblastoma is also able to spread outside the CNS, but rarely does so[9]. However, these symptoms are not sufficient to diagnose MB. In the vast majority of cases MB is considered an important differential diagnosis after initial radiological imaging. Thereby, the space-occupying lesion in the posterior fossa is identified and then removed surgically. Following tumor resection or biopsy, the removed tumor tissue is used to establish an MB diagnosis[9,30]. MB is not clearly distinguishable from Pilocytic Astrocytoma (PA), Ependymoma (EPN) and other pediatric brain tumors in radiography pictures of the patient, hence tissue-based histological classification is required and often complemented by molecular diagnostics. In the clinic, MB is diagnosed by neuropathological exclusion of other cancer types, such as atypical teratoid/rhabdoid tumor (AT/RT), glioblastoma, ependymoma and others[9].

Risk stratification for MB patients is not only based on the neuropathological diagnosis, but also relies on metastatic status and age of the patient[30]. Today, MB groups are being distinguished by molecular analyses, facilitating clinical risk stratification[10,23,31]. The common treatment approach for MB involves surgical removal of the tumor[23], risk stratification of the patient, followed by radiotherapy and chemotherapy adapted to risk and patient age[30]. The overall 5-year survival of MB patients is around 70%, however, a significant proportion of these patients experience long-term side-effects due to surgery, radiotherapy and chemotherapy[7,30,32]. Group assignment is commonly conducted using histopathological assessment, immunohistochemistry, DNA methylation profiling, or RNA-sequencing (RNA-Seq)[18,26,33–36]. Distinction between WNT MBs and SHH MBs is straight forward, however, several studies demonstrated a significant variability in the diagnostic distinction between Group 3 and Group 4 MB[17,33]. Currently, DNA methylation profiling is considered the gold standard for MB group classification[9].

1.2.3 Epidemiology

Medulloblastoma is the most common malignant brain tumor in children, with an estimated annual incidence of approximately five newly diagnosed patients per 1 million children and adolescents[9]. Medulloblastoma is most commonly diagnosed in young children (incidence peaking at ~6 years of the patients age), but also occurs in adults (~1% of MB cases). Medulloblastoma is found 1.8 times more often in males than in females, but this also depends on the MB group[9,32]. Ethnicity does not play a significant role in survival[32].

1.2.4 WNT medulloblastoma

WNT MB is the molecular group with the best overall survival for patients, yet it only comprises ~10% of all MB cases[18,25,26,32,37]. Recurrent genomic alterations in WNT MB are found in *CTNNB1* and *SMARCA4,* among others[23], and a monosomy of chromosome 6[18,23] is a hallmark feature. The *CTNNB1* mutations that typically occur in WNT MBs result in a constitutively active beta-catenin that is part of the WNT signaling pathway[38], making this pathway permanently active and thereby promoting tumor growth. *TP53* mutations can also be detected in WNT MBs, but this does not indicate better or worse prognosis for these patients[31]. WNT MB presents with the lowest frequency of metastatic disease at time of diagnosis when compared to the other MB groups. The histology of these tumors is mostly classic. Since the outcome of WNT MB patients is - if treated accordingly - favorable, clinical studies aim to de-escalate therapy in order to minimize treatment sequelae while still trying to achieve long-term survival. The subtypes of WNT MBs are alpha and beta, with alpha tumors being more common in younger patients and beta tumors showing less frequently chromosome 6 monosomy. The typical location of WNT MBs is in the brainstem, and the sex ratio of WNT MB patients is 1:1 [23]. Germline WNT pathway mutations may lead to Turcot syndrome[25]. The 5-year survival for patients with WNT alpha and beta MBs is 97% and 100%, respectively. Primarily, WNT MB mortality can be attributed to complications of the therapy or secondary neoplasms[9]. The proposed cellular origin of WNT MBs are progenitor cells in the lower rhombic limp[23].

1.2.5 SHH medulloblastoma

Sonic Hedgehog (SHH) MBs comprise ~ 30% of all MB diagnoses. Common genomic alterations are found in SHH pathway genes, such as *PTCH1*, *SMO*, *GLI1/2*, *MYCN*, and *SUFU*, but also genes that correspond to syndromes of genome instability such as *TP53*, which is linked to Li-Fraumeni-Syndrome[39]. With generally lower survival compared to WNT MBs, SHH MBs are divided into four distinct subtypes: alpha, beta, gamma and delta[18]. Beta and gamma SHH MBs are more common in infants and frequently show desmoplastic or MB with extensive nodularity (MBEN) histology. Beta SHH MBs are associated with the worst 5-year survival rate of all SHH subtypes with ~ 67%, and often carry a *PTEN* gene loss. Beta SHH MB patients also face a three-fold increase in the chance of having metastatic dissemination at diagnosis compared patients with gamma SHH MB (33% in beta compared to 9.4% in gamma SHH MBs), whereas SHH MB gamma survival rates are more favorable at 88%, and these tumors typically show a balanced genome. The other two SHH MB subtypes, alpha and delta, are more common in older children. Alpha SHH MBs show typical SHH pathway alterations, such as amplifications of *GLI2* and *MYCN*, in addition to *TP53* mutations. Unlike WNT MBs, *TP53* mutations do have an impact on survival, reducing the patient survival rates by 50%[31]. Additionally, SHH MB alpha is characterized by a loss of chromosome arms 9q, 10q and/or 17p. Notably, the delta SHH MBs comprise the SHH subtype with the highest 5-year survival, at 88.5%, and this subtype frequently harbors *TERT* promoter mutations[18]. The typical anatomic location for SHH MB is the cerebellar hemisphere and the cell type of origin likely is the granule cell precursor cell[23].

1.2.6 Group 3 medulloblastoma

The overall prognosis for Group 3 MB remains poor. Group 3 MB patients comprise 25 % of diagnosed MB patients, and metastasis are more common among these patients compared to all other MB groups. Dependent on the molecular subtype, metastesis are seen in 43.4% (alpha), 20%(beta) and 39% (gamma) of patients[18]. *MYC* amplification is most commonly detected in Group 3 gamma, and this subtype is associated with a particularly poor survival rate compared to all other MB subtypes. Group 3 beta MBs show overall genetic instability characterized by broad genomic losses and gains of whole chromosome regions such as chromosome 7 and/or 18 gain[23], isochromosome 17q and common losses of chromosomes 8, 11, 10q and 16q. *OTX2* amplification and *DDX31* loss are enriched in this subtype that is also frequently associated with high *GFI1/1B* expression. Metastatic dissemination at diagnosis is most commonly detected in Group 3 alpha MBs. However, this subtype is associated with a relatively favorable 5-year survival rate (66.2%) compared to other Group 3 subtypes. Overall Group 3 MBs are commonly characterized by classic or anaplastic/ large-cell histology and a 2:1 male:female sex ratio[25,37]. Neural stem cells located at the fourth ventricle are currently considered as the likely cellular origin[23].

1.2.7 Group 4 medulloblastoma

Group 4 MB is generally less aggressive as compared to Group 3 MB, but is still associated with an intermediate prognosis. Group 4 MB comprise ~ 35% of all MB diagnoses, and this type is more frequently observed in males than in females, at a ratio of 3:1. Group 4 MBs are subdivided into three subtypes: alpha, beta and gamma. Isochromosome 17q constitutes the most common alteration in Group 4 MBs. Alterations with subtype-specific enrichment include *MYCN/CDK6* amplification in Group 4 alpha, *SNCAIP* duplication in Group 4 beta, and *CDK6* amplification in Group 4 gamma samples. Metastases are seen in approximately 40% of patients across these three subtypes. The 5-year survival rates differ significantly across these Group 4 subtypes with 66.8%, 75.4%, and 82.5%, in alpha, beta, and gamma subtypes, respectively. Unipolar brush cells are considered the likely cell type of origin[18,23] .

Notably, Group 3 and Group 4 MB show a distinct protein/RNA imbalance[33(p4)], regarding the predominant level of oncogenic pathway dysregulation. Specifically, this enrichment is observed at the transcriptional or proteomic levels in Group 3 and Group 4 MBs, respectively. This discrepancy is not detected in WNT or SHH MB, and a mechanistic explanation is still missing[33(p4)].

1.2.8 Medulloblastoma risk stratification

Risk stratification for MB patients is based on several clinical and biological properties of the tumor[20,37,40]. Despite steadily increasing survival rates, there is still a pressing clinical need for biomarkers that are fast, cheap, easy and reliable to classify MB groups and improve risk stratification. The overall goal here is to make the therapeutic intervention as effective as possible, while side effects and neurocognitive sequelae should be minimized[33,34,41]. Examples for important MB biomarkers include WNT and SHH groups. WNT MB can be identified by immunohistochemistry[20], but also recurrent genetic alterations, such as monosomy 6. For MB as in many cancer types, *MYC* gene amplification is a biomarker for poor prognosis[20,42]. The distinction between Group 4 (intermediate prognosis) and Group 3 (worse prognosis, *MYC* amplification) MBs is sometimes not precise, and thus there are still efforts required to define reliable biomarkers for diagnostic distinction of the two groups[33]. Furthermore, MB group identification is important to select targeted therapies for molecularly defined patient cohorts including *SMO* inhibitors in SHH MB[16].These examples provide compelling rationale why accurate molecular classification is urgently required to improve outcome of MB patients. Integration of subtype classification comprises the next step in this even more specific treatment optimization – ultimately aiming for personalized medicine approaches, where each patient has a therapy tailored based on diagnosis and omics data[27].

1.3 RNA sequencing

To this date, several RNA quantification methods have evolved enabling the precise detection and quantification of several RNA types including long non-coding RNA (lncRNA), microRNA and mRNA[43,44]. The general RNA-Seq workflow includes a first isolation of RNA from the sample to be investigated. Trizol-based RNA isolation was the default method for several years, while alternative approaches for RNA extraction are now increasingly used, including new, modern approaches[45]. The RNA molecule is rather instable. Therefore, the isolated RNA needs to be converted to its complementary DNA (cDNA) that is a more stable molecule currently used for next-generation sequencing approaches. The cDNA is fragmented in a next step, an adenosine overhang is added to each cDNA fragment on which the library adapters are ligated to in a subsequent step. In a currently available Illumina sequencing machine such as the HiSeq2500, these libraries are then added to the sequencing flow cell and amplified to form a cluster of single-stranded cDNA copies originating from the same original cDNA molecule. The sequencing by synthesis is generally a repetition of 3 steps executed 100-250 (depending on the machine and the chosen workflow) times: extending the DNA by one base that includes a fluorescent label that differs for each of the four DNA bases, reading the optical signal indicating which base was incorporated and deblocking of the next synthesis step. This happens on the complementary strand, resulting in a paired DNA molecule in the end. For paired end sequencing, the synthesis is made from both directions, for single-end reads this is only the case for one side[46]. The sequencing machine itself puts out base calls, or rather a record of what colors were detected where on the flow cell at what cycle. These base call (bcl) files are then used for subsequent computational analysis.

During the demultiplexing step, each read is sorted into sample-specific files. These files are in the .fastq format, including the sequences of RNA-Seq reads, quality parameters of the read, and other meta-data such as information about the utilized flowcell and sequencing machine as well as the time of the run.

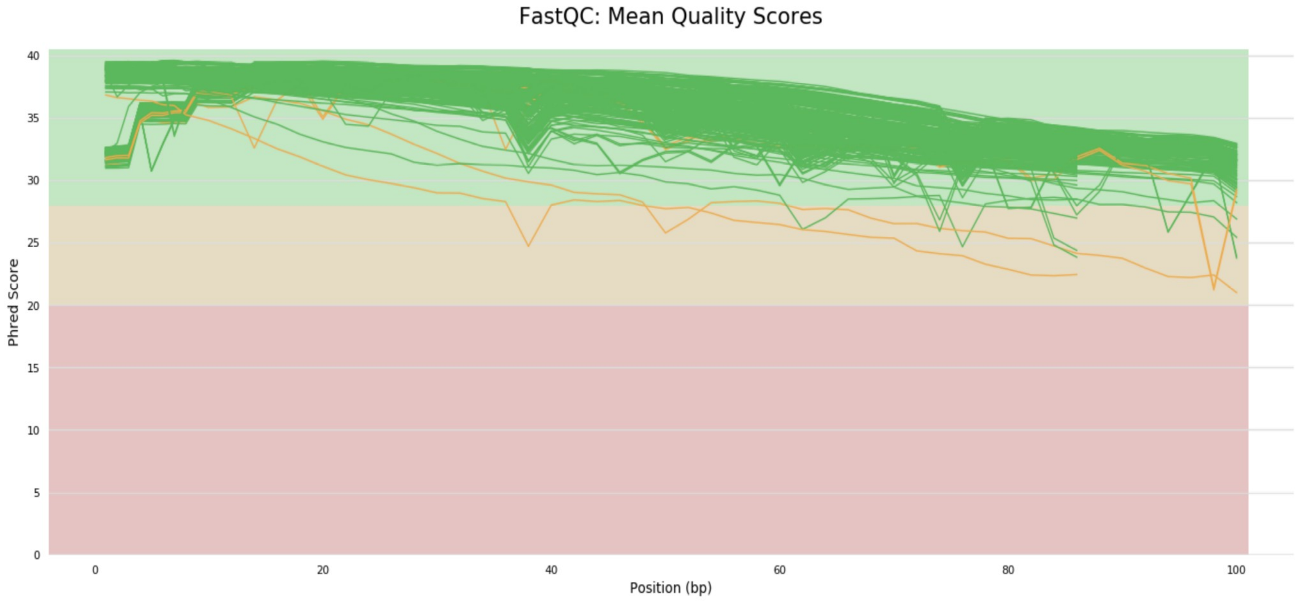


Figure 2: FastQC report summarized by MultiQC. Each horizontal line represents one sample and the mean quality (Phred) score in the corresponding position. The green area indicates generally good quality data, yellow acceptable and red bad quality. At least the two fastq files indicated by the yellow horizontal lines should to be discarded.

The resulting RNA-Seq reads are filtered based on the quality and the length of the read (Figure 2) including adapter contamination (Figure 3). These parameters can be assessed and visualized with multiple tools. Figure 2 shows a MultiQC example output graph with multiple RNA-Seq samples as input. Two samples with inferior RNA-Seq quality are identified (yellow lines). The filtered RNA-Seq reads can then be aligned to a reference genome utilizing gap-aware mappers such as STAR[44,47,48]. The following RNA-Seq quantifications need to be normalized and can subsequently be used to quantify RNA expression , search for single nucleotide variants (SNVs), and to detect circRNAs[49–51].

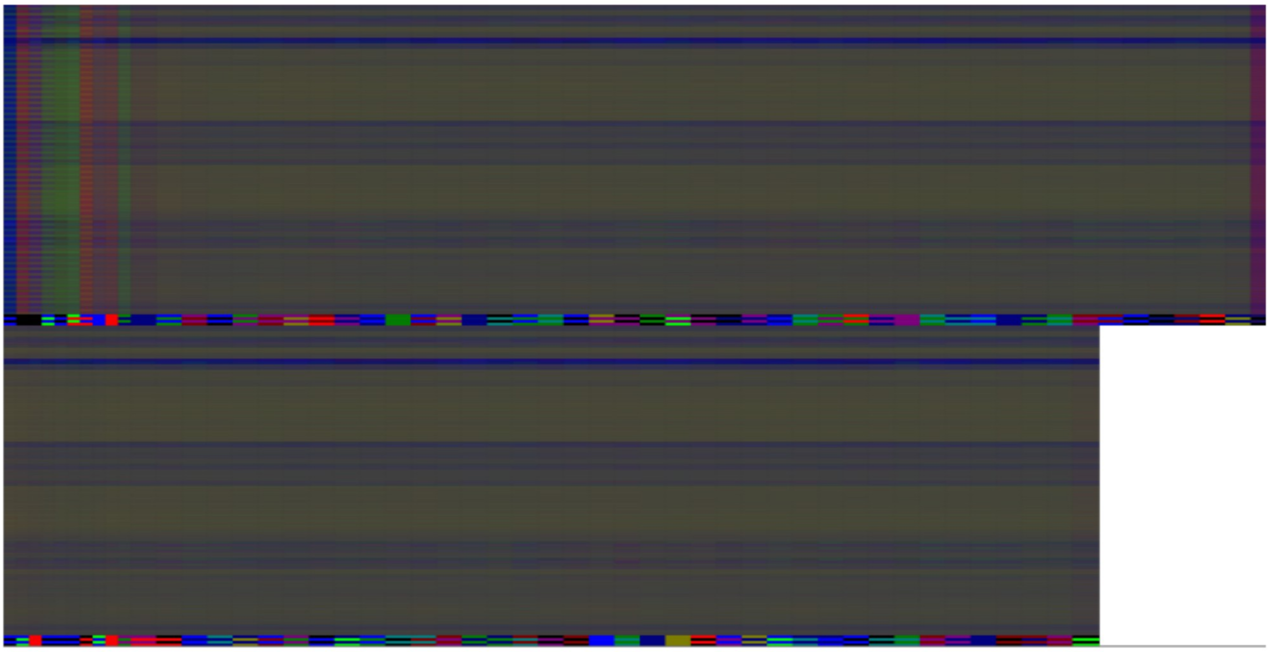


Figure 3: Per base sequence content plot from the same MultiQC report. Here the first few bases depict a clear shift into one of the four colors, indicating bad quality or adapter contamination. This adapter content will be removed before the analysis.

1.4 Circular RNA

1.4.0 Overview

Circular RNA is a closed loop of single-stranded mostly non-coding RNA without poly(A) tail and cap structure[52–54]. In contrast to linear RNAs, this kind of non-coding transcript is RNAseR resistant[49] and characterized by a longer half life time[55]. Historically, circRNA transcripts have been discarded[54] as junk but due to recent improvements in RNA-Seq data quality and quantity it is possible now to quantify circular RNAs reliably using different methodologies[56]. The expression of circRNAs has been shown to be developmental stage, tissue- and time-specific[57]. Furthermore, circRNA expression is enriched in human neuronal tissue[57–61]. Some circRNAs show a high degree of conservation, and seem to be present in the whole eukaryotic tree of life[62] leading to a circRNA-based theory of the origins of first functional biomolecules[63]. Another focus of circRNA research is based on viruses, where virus-encoded circRNAs have been detected[64–67]. An *in vivo* process for linearization of circRNA - its breakage of the ring structure to form a linear transcript - has not yet been described.

1.4.1 Classification of circular RNAs

Based on their genomic content, circRNAs are classified into exclusively exonic circRNAs (EcircRNAs), intronic and exonic circRNAs (EIcircRNAs)[68] and intronic lariats escaping degradation, becoming intronic circRNAs (IcircRNA)[69,70]. Exonic circRNAs include a subclass of annotated start-codon including AUGcircRNAs, which are more evolutionary conserved and are synthesized in a mostly ALU-independent manner[71].

1.4.2 Synthesis of circular RNAs

In eukaryotes, circRNAs are synthesized from pre-mRNA *in vivo* by back-splicing. Back-splicing involves a splicing donor and a acceptor site that are ligated together, forming a closed ring of RNA[56]. This process can be aided by reverse complementary sequences in close genomic proximity to the two splice sites of the circRNA (ALU[72,73], a primate-specific genomic element[74]) or an enzymatic binding site facilitating the back-splicing (e.g by QKI[75]), while ADAR1 binding sites antagonize circularization in proximity. Tissue-specific circRNA synthesis regulators have been found as well for neuronal tissue[76]. Most circRNAs are synthesized by joining two canonical splice sites[77], namely AG-GU[78], but there are exceptions especially for IcircRNAs[70]. Another circRNA-forming mechanism includes exon-containing lariat precursors that originated in exon-skipping events during splicing[73]. The RNA-interacting protein SFPQ can also have effects on circRNA biogenesis[74]. While being synthesized, circRNAs themselves can be alternatively spliced, pointing towards a diverse circRNOME in each cell[79]. If the function of the spliceosomal machinery is limited, the output of coding genes can be shifted to increased circRNA formation[80], especially in genes where the circRNA transcript is in competition with the linear transcript.

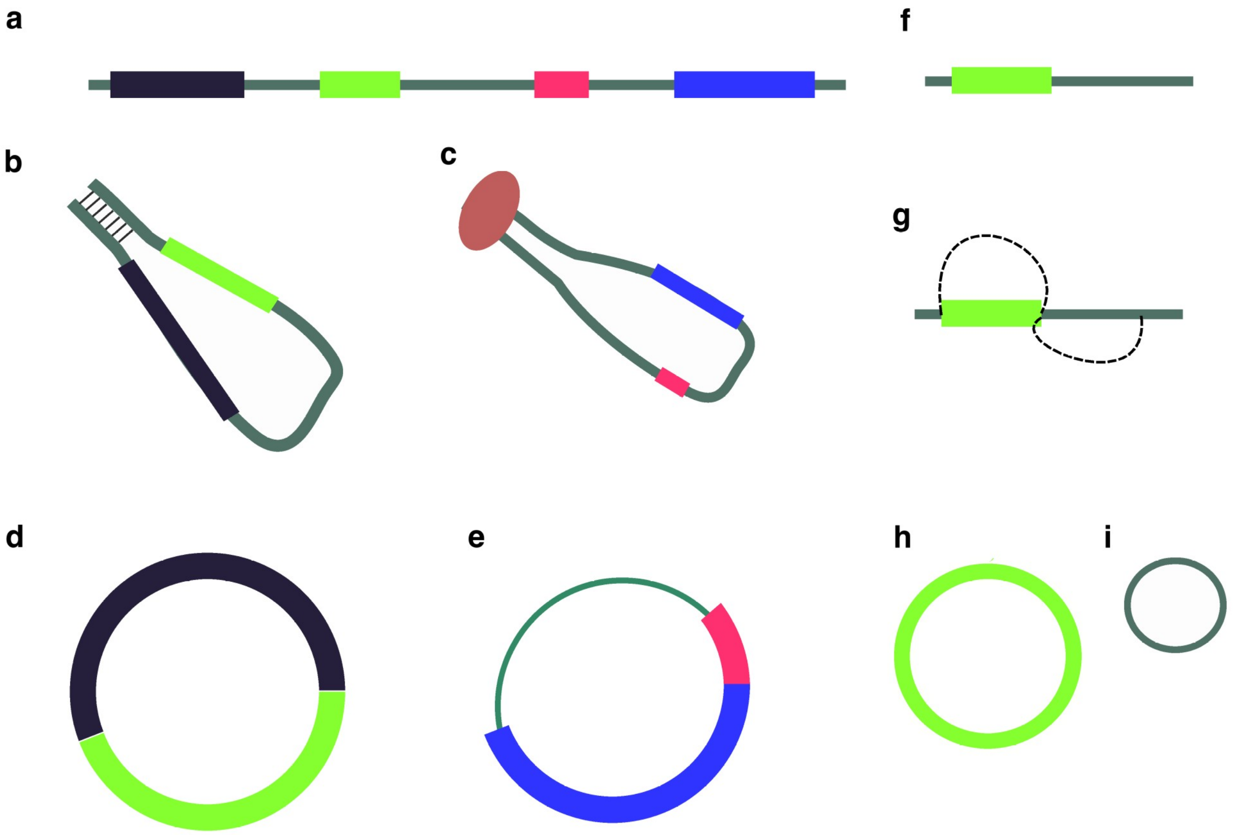


Figure 4: Overview of circular RNA synthesis pathways. a) Linear DNA with introns and several exons. b) Back-splicing by complementary sequences flanking the exons. c) RNA binding protein (RBP) aided back-splicing. d) Exonic circular RNA with 2 exons. e) EIcircRNA with 2 exons and 1 intronic region. f) DNA with 1 exonic and 1 intronic region. g) Single-exon circular RNA synthesis by back-splicing and intron lariat formation. h) Single exonic circular RNA. i) Intronic lariat.

1.4.3 Functions of circular RNAs

The most prominent putative function of circRNAs is their ability to sponge microRNAs[81,81,82]. Other functions that have been described are the sponging of RNA binding proteins (RBPs)[49]. Most circRNAs are non-coding, although some circRNAs include an internal ribosomal entry site (IRES) and have been shown to be actively translated into proteins[83,84]. There is another putative function of circRNAs that has been discovered: formation of protein scaffolds[85]. Exonic intronic circRNAs are most commonly linked to this function. Overall, the function of most circRNAs is unknown to this date and still needs to be elucidated.

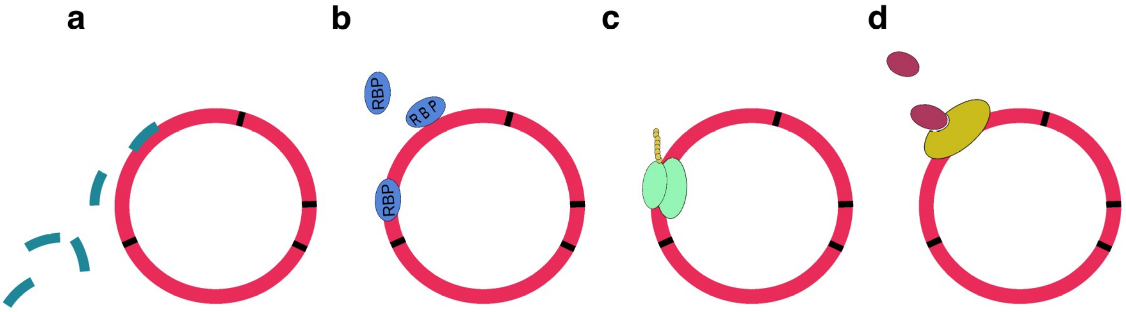


Figure 5: Examples of circular RNA functions. a) MicroRNA sponging. b) RNA binding protein RBP sponging. c) Translation of a circRNA into a protein. d) Complex forming including a protein and a cofactor.

1.4.4 Circular RNAs as disease-specific biomarker

Circular RNAs have multiple properties that make them ideally suited for the use as biomarkers. They are remarkably specific regarding species-, age-, developmental-stage- and tissue-dependent expression, suggesting a high degree of differentially expressed circRNAs that could correlate to various properties of the disease[86]. Multiple *in silico* detection methods are available for different raw data inputs and RNA-Seq protocols. Once candidate biomarkers are identified, polymerase chain reaction (PCR) based approaches, such as quantitative real-time PCR (qRT-PCR), can be used to detect these in a fast and an inexpensive manner.

Cancer has been one of the most investigated research fields for circRNA biomarker studies[73,86–89], including hematological maligancies[89]. Since circRNAs are especially enriched in neuronal tissue[57], more stable than linear RNAs and detectable by qRT-PCR of selected back-splice junctions[49,59–61], circRNAs have been investigated in several brain cancer entities for their potential as biomarker[82,87,90]. Different circRNAs and their respective deregulation have been associated with several hallmarks of cancer, suggesting a role in tumorgenesis[73]. Furthermore, circRNAs hold great potential for liquid biopsies including detection in cerebro-spinal fluid (CSF), which might allow for clinical applications in brain tumor patients including disease monitoring.

Other ways of using circRNA as biomarker for cancer is exemplified by Okholm et. al.([91]), who utilized the overall abundance of circRNA as an indicator and not single circRNA transcripts. Furthermore, circular RNAs have been shown to also exist in exosomes[92–94] and are even enriched in some biofluids compared to their tissues of origin in the human body[95]. Hence, circRNA detection in biofluids holds great promise as a convenient and non-invasive direction for biomarker development.

1.5 Detection of circular RNAs

1.5.0 Overview

Electron microscopy[96] was first used to identify circRNAs, but since then several other identification techniques have evolved. CircRNA can be detected by many different methods, *in silico* and *in vitro.* The circRNA detection *in silico* relies on analysis of RNA-Seq data[96]. *In vitro* assays such as nanoString, Northern blot, droplet digital PCR (ddPCR) or qRT-PCR with primers detecting the back-splice junction of the circular transcript are relatively low-throughput in comparison[25]. Another approach is the circleseq protocol[72]. Here, the sample is treated with RNAseR before sequencing, exploiting circRNAs stability toward this enzyme and thus enriching for circular RNAs while depleting linear transcripts. This method can also be used to confirm circRNAs found in other datasets, although some circRNAs have been shown to be RNAseR sensitive[49].

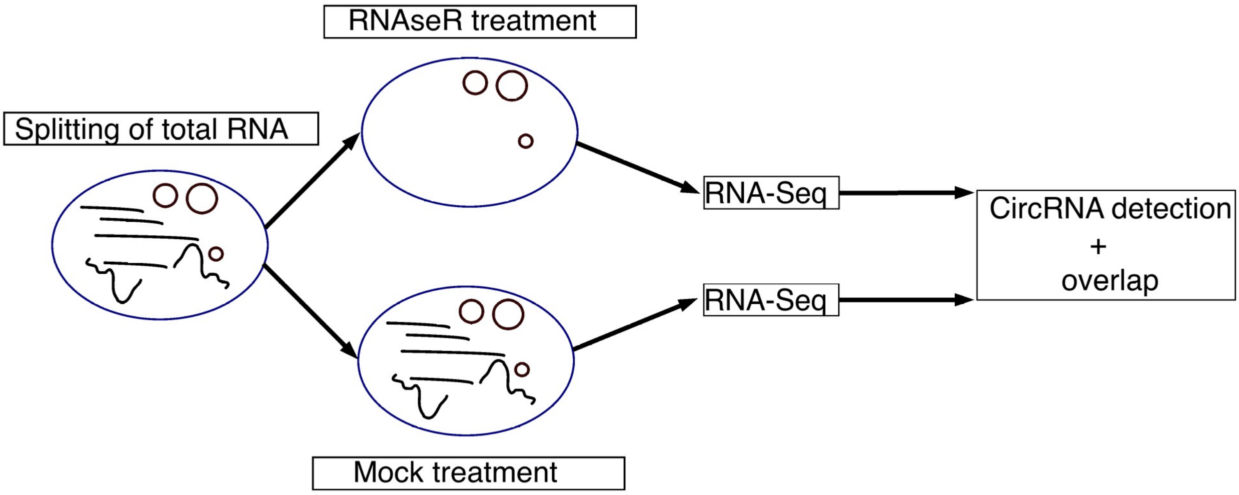


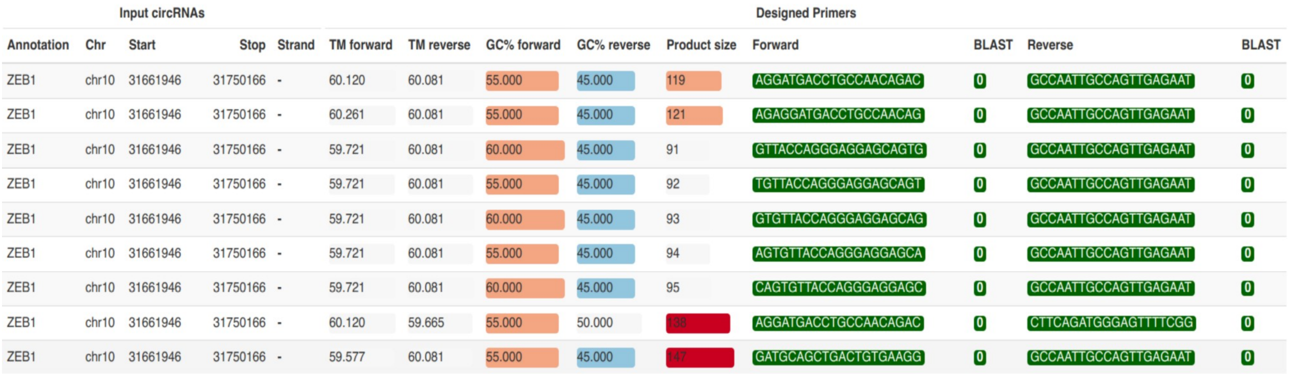
Figure 6: The workflow of a typical CircleSeq protocol. The sample is split into two parts after RNA isolation and ribosomal RNA depletion. One of the two preparations is then treated with RNAseR while the other preparation is not. After sequencing both samples RNAseR- stable (circular RNAs that are enriched in the RNAseR treated sample) circular RNAs can be identified to confirm the presence of circRNAs in the untreated preparation.

1.5.1 Detection of circular RNAs *in silico* based on RNA sequencing data

Multiple *in silico* circRNA detection methods have been developed to analyze RNA-Seq data and identify circular transcripts[53,96,98–100], all relying on back-splice junction detection. The linearly spliced parts of any circRNA can originate from linear or circular transcripts, so reads with back-splice junctions are the only circRNA-specific signal in this context. However, back-splice junction reads are no guarantee of definitive circRNA detection, since several genomic parts get “mixed” in a process called exon scrambling[101], during which back-splice junctions are formed as well, but do not result in circRNA formation. Since these *in silico* methods alone have a high false-positive rate[49], it was proposed to use multiple detection pipelines simultaneously, and to accept only repeatedly detected circRNA transcripts as true positives[102]. This approach has been applied by several research teams[103,104].

1.5.2 Detection of circular RNAs *in vitro / in vivo*

CircRNAs can be detected by PCR of the back-splice junction, but this is a comparatively low throughput method, as the detection assay needs to be designed individually and measurements are done for each single circRNA and each sample separately. The primers designed for each circRNA need to be placed in close proximity to the back-splice junction in order for it to be covered by the PCR product. Circtools includes circtools primer (see Figure 7), a tool for designing primers for this application[105].



Input circRNAs					Designed Primers								
Annotation	Chr	Start	Stop	Strand	TM forward	TM reverse	GC% forward	GC% reverse	Product size	Forward	BLAST	Reverse	BLAST
ZEB1	chr10	31661946	31750166	-	60.120	60.081	55.000	45.000	119	AGGATGACCTGCCAACAGAC	0	GCCAATTGCCAGTTGAGAAT	0
ZEB1	chr10	31661946	31750166	-	60.261	60.081	55.000	45.000	121	AGAGGATGACCTGCCAACAG	0	GCCAATTGCCAGTTGAGAAT	0
ZEB1	chr10	31661946	31750166	-	59.721	60.081	60.000	45.000	91	GTTACCAGGGAGGAGCAGTG	0	GCCAATTGCCAGTTGAGAAT	0
ZEB1	chr10	31661946	31750166	-	59.721	60.081	55.000	45.000	92	TGTTACCAGGGAGGAGCAGT	0	GCCAATTGCCAGTTGAGAAT	0
ZEB1	chr10	31661946	31750166	-	59.721	60.081	60.000	45.000	93	GTGTTACCAGGGAGGAGCAG	0	GCCAATTGCCAGTTGAGAAT	0
ZEB1	chr10	31661946	31750166	-	59.721	60.081	55.000	45.000	94	AGTGTTACCAGGGAGGAGCA	0	GCCAATTGCCAGTTGAGAAT	0
ZEB1	chr10	31661946	31750166	-	59.721	60.081	60.000	45.000	95	CAGTGTTACCAGGGAGGAGC	0	GCCAATTGCCAGTTGAGAAT	0
ZEB1	chr10	31661946	31750166	-	60.120	59.665	55.000	50.000	138	AGGATGACCTGCCAACAGAC	0	CTTCAGATGGGAGTTTTCGG	0
ZEB1	chr10	31661946	31750166	-	59.577	60.081	55.000	45.000	147	GATGCAGCTGACTGTGAAGG	0	GCCAATTGCCAGTTGAGAAT	0

Figure 7: Output of a circtools primer run. For each circular RNA, multiple potential primers are designed, giving the user a choice. Forward and reverse columns show the primer sequence, basic local alignment search tool (BLAST) reveals the uniqueness in the genome. The GC% forward and reverse depict the Guanine and Cytosine (GC) content of the primers shown.

Rolling circle amplification[106] is another method to detect circRNAs, especially when the circular-to-linear ratio is expected to be low (giving a high amount of noise). Microarray techniques for circRNA are available since 2014[96], making circRNA detection even more accessible. CircRNAs can also be detected by the circSCREEN method, utilizing live-cell imaging and specifically engineered circRNAs that express GFP proteins upon circularization[75].

1.6 Circular RNA and medulloblastoma

The circRNA landscape of MB was previously not comprehensively studied in sufficiently sized primary tumor cohorts. The discovery of circRNA-based biomarkers holds great promise due to the molecular heterogeneity of the disease and highly distinct outcomes to current treatment approaches of MB patients. Previous circRNA studies in MB[90] did not aim to resolve the whole MB group circRNA landscape. Other brain cancers have been a target of exploratory circRNA biomarker studies before[82,104,107] with several candidate circRNA biomarkers that were discovered. As MB is usually treated by removing the tumor mass surgically, tissue becomes available not only for pathological but also biological and bioinformatic analysis. The CSF could be an additional source for circRNA biomarker discovery as described above.

1.7 Other tools for circular RNA research

In the years following the first publication of circRNA detection[72] algorithms several *in silico* circRNA tools emerged[96,49,53,56,98–100,102,103,118–123]. This landscape of circRNA tools includes databases of circRNAs[88,95,108–110,110,111], circRNA interaction and sequence analysis tools[105,108,112–115], statistical analysis of circRNA data[105] as well as tools for visual circRNA representation[116,117]. A plethora of circRNA detection algorithms was also developed.

1.8 Snakemake

Bioinformatic workflows often combine a multitude of programming languages, software modules and tools utilized. This poses a risk: when a bioinformatic workflow or parts of it are updated, final results may change. This is a problem to the reproducibility of such pipelines. One update of any small software component may result in different final findings.

Snakemake[125] is a framework for reproducible data analysis not only delivering the ability to deliver stable software environments for each separate step of a data analysis pipeline via conda (https://docs.conda.io/en/latest/), but also the possilility to orchestrate many small substeps of complex pipelines on computing clusters with many nodes and one local machine. Snakemake is a rule-based bioinformatic workflow management tool that can be used to execute commands based on rules. These rules usually include an input file or a list of input files, an output file or a list of output files, a software environment in which the commands included in the rule are executed, and one or multiple commands that convert the input files into output files. There are several options to add parameters to each rule, such as the number of CPU threads that can be used. Rules are connected by files: output files of one rule are input files for the next rule, resulting in the here called rule “all” that collects all final output files of the workflow.

Each shell command can be wrapped into a rule with independent software environments. Snakemake offers rich documentation, several tutorials and a “snakemake workflow catalog” at https://snakemake.github.io/snakemake-workflow-catalog/, where several snakemake-based pipelines are listed.

2 Aim of this thesis

Medulloblastoma is the most common malignant brain tumor in childhood. The disease-related morbidity is high and side-effects of its multimodal therapies may result in major burdens for survivors of the disease. Refined risk stratification and classification of the disease are urgently needed to tailor therapy intensity to the individual risk profile of the patient. The identification of the biological subgroups, WNT, SHH, Group 3 and Group 4, has led to the recognition of significant intertumoral heterogeneity with important biological, clinical and prognostic associations. Hence, reliable biomarkers for refined biological classification of the disease are urgently needed. The value of mining non-coding RNA for biomarker discovery in MB research still remains unclear. In particular, the circRNA landscapes of the distinct MB groups have not yet been explored comprehensively. Notably, circRNA profiling has shown a remarkable potential for biomarker development, and thus constitutes a promising tool for therapy and risk stratification in several malignancies.

The aim of this study was to evaluate circRNA profiles and define reliable circRNA-based biomarkers for MB groups in order to refine existing classification approaches to the disease.

To determine the biomarker potential of circRNAs in MB, the first goal was the establishment of a multi-pipeline workflow that was reliable and precise, leveraging the full potential of this layer of information. Next, a discovery dataset was evaluated for circRNA expression profiles to identify sets of biomarkers that, in the best case, show highly differential expression patterns correlating to clinical features, metastatic dissemination, prognosis, or other features such as molecular MB subgroups. Candidate biomarkers were then validated in a non-overlapping, independent validation cohort and biomarker specificity was determined in comparison to normal and other cancer tissues. Once reliable and specific circRNA MB biomarkers were defined, the candidates were verified using targeted, orthogonal experimental approaches. With a novel, accurate circRNA-detection pipeline, and utilizing unparalleled amounts of circRNA MB data, we aimed to identify biomarkers that will lead the way to fast, inexpensive and robust MB group classification in order to improve patient stratification of children and adolescents with this highly aggressive disease in the future.

3 Materials and methods

3.1 Wet lab experiments

3.1.1 Patient samples

All MB samples were collected following written informed consent. Approval for the study was given by the internal review board at the Necker Hospital for Sick Children (Paris, France, IRB approved protocol number DC-2009-955 for tumor banking) and by the internal review board of the Medical Faculty at Heinrich Heine University Düsseldorf (study numbers 3005 and 2018-45-FmB). All analyzed samples were collected from newly diagnosed MB. RNA-Seq data from healthy fetal brain tissue was downloaded from ENCODE (https://www.encodeproject.org/). From each sample included in the discovery (EGA: EGAD00001004327) or the validation cohort, total RNA was prepared and sequenced as described elsewhere[18]. The validation dataset is not public as of 03/11/2022.

3.1.2 *MYC*/RNAseR cell line models

UW228, DAOY and ONS76 cell lines were included in the *MYC*/RNAseR dataset. All cell lines were cultured in DMEM medium with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin (P/S) at 37°C in a 5% CO_2 atmosphere. Cell line authenticity was proven by short tandem repeats (STR) profiling, contamination based on mycoplasma was ruled out by PCR testing. Cell culture work was kindly performed by Sarah Göbbels.

Table 1: Consumables used for medulloblastoma cell line cultivation.

Consumable	Catalog ID	Producer
DMEM Medium	31966-021	Thermo Fischer Scientific
FBS	F9665/ P30-3302	Sigma-Aldrich/ Pan Biotech
P/S	P4333	Sigma-Aldrich

3.1.3 Lentiviral vector construction for *MYC* overexpression

Stable *MYC* overexpression was achieved by lentiviral transduction. The vector used was previously constructed by Dr. Viktoria Marquardt based on the LeGO-iG2 (Addgene ID 27314) vector ligated with the cDNA sequence of *MYC* derived from the pcDNA3.3_c-*MYC* plasmid (derived from Addgene ID 26818). Successfully transduced cells were identified and filtered by flow cytometry for GFP expression utilizing the MoFLo XDP (Beckman Coulter). This experimental work was kindly performed by Dr. Nan Qin.

3.1.4 *MYC*/RNAseR RNA isolation and sample preparation

RNA was collected using the Maxwell RSC instrument with the manufacturers' RNA isolation kit and a RNA Integritiy Number (RIN) of >9 was confirmed for each sample not treated with RNAseR. A total of 100ng of isolated RNA of each sample was used for reverse transcription and library preparation (according to Illumina, USA, low throughput protocol, ID RS-122-2001). Libraries were validated and quantified using DNA1000 and high-sensitivity chips on Bioanalyzer. 7.5pM of the denatured libraries were used as cBot input (Illumina, USA). After this, the prepared samples were sequenced on an Illumina HiSeq2500 machine (Illumina, USA). RNA collection and sample preparation were kindly performed by Frauke-Dorothee Meyer.

3.2 Dry lab experiments

In the following paragraphs the software and hardware packages used in this thesis are listed. Additional custom code can be acquired from Github (https://github.com/daaaaande), GitLab (https://gitlab.com/daaaaande/circs) or upon request from danielrickert@protonmail.com. All data analysis and code writing was performed by Daniel Rickert.

3.2.1 Software packages

Major software packages used in this thesis are listed below in Table 2.

Table 2: software packages used in this thesis.

Software package	**Source**
R	https://www.r-project.org/
Inkscape	https://inkscape.org/de/
STAR	https://github.com/alexdobin/STAR
Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
find_circ	http://circbase.org/cgi-bin/downloads.cgi
DCC	https://github.com/dieterich-lab/DCC
CIRCexplorer	https://github.com/YangLab/CIRCexplorer
bedtools	https://sourceforge.net/projects/bedtools
Samtools	http://samtools.sourceforge.net/
Snakemake	https://github.com/snakemake/snakemake
Perl	https://www.perl.org/get.html
FastQC	https://sourceforge.net/projects/fastqc.mirror/
MultiQC	https://multiqc.info/
RStudio	https://www.rstudio.com/products/rstudio/download/

3.2.2 Non-default Perl packages

The non-default Perl packages used in this thesis are listed in Table 3.

Table 3: Non-default Perl packages used during this project.

Perl Package	**Purpose**
Parallel::ForkManager	Multi-threading of matrixmaker_V4.pl
Getopt::Long	Input parameter parsing

3.2.3 Non-default R packages

The non-default R packages used in this thesis are listed below in Table 4.

Table 4: Non-default R packages used in this thesis.

R package	**Purpose**

dplyr	Data subsetting
VennDiagram	Voting visualization

3.2.4 Hardware used

The hardware used during this thesis is listed in Table 5 below.

Table 5: Hardware utilized during this thesis.

Hardware	**Primary purpose**
HPC Computing Cluster HILBERT	CircRNA detection pipeline
Thinkpad L480	Data analysis
TERRA-Server	Data repository

3.2.5 RNA sequencing pre-processing

Raw RNA sequencing data were demultiplexed based on unique adapter sequences and converted to fastq format using the bcl2fastq package, ensuring adapter sequences were not masked. Sequencing read quality was then assessed using the FastQC sequencing quality control package and summarized with MultiQC. All samples passed all quality controls.

3.2.6 Circular RNA detection using the *circs* workflow

Circular RNAs were quantified with a three pipelines workflow including DCC, find_circ and CIRCexplorer1. We named this analytical pipeline *circs*. Briefly, *circs* used custom written Perl scripts to locally execute DCC, find_circ and CIRCexplorer1 (the automated version includes a non-default choice of STAR as the aligner) pipelines, overlapped the data to minimize false positives and normalizes the final data output. First, each circRNA detection pipeline was run according to the input data format (here both datasets were paired-end). Second, each of the three circRNA output datasets was filtered to include only circRNAs that are detected with at least two junction reads in at least one sample. Third, voting includes the overlap of the three filtered output tables: Only circRNAs that were present in all three filtered output tables were accepted for further analysis. In a last step, the DCC quantifications of accepted circRNAs were normalized to DCC circRNA backsplice junction reads per million total RNA sequencing reads to ensure comparability across data sets. The normalized, filtered and voted DCC circRNA quantifications were then used for downstream analysis.

3.2.7 Data processing

All downstream analyses were performed in R. For clustering, the data were first normalized, filtered for the top 500 differentially expressed (highest standard deviation across all samples) circRNAs, and then clustered according to Pearson's dissimilarity (1-(average Pearson's correlation). For statistical comparison of circRNA expression across MB groups, ANOVA was used with Tukey’s HSD as post-hoc test. All p-values shown in this document are adjusted p-values if applicable.

3.2.8 MiOncoCircDB overlap and comparisons

For MiOncoCircDB overlaps of all three datasets (discovery, validation and healthy samples), the liftOver tool was used to convert the genomic coordinates of all three datasets from hg19 into hg38. In case that the respective coordinates could not be correctly annotated, circRNA coordinates were excluded from all following analysis steps. To compare *circs* data to MiOncoCircDB expression data, *circs* CIRCexplorer1 output was normalized to the median number of RNA sequencing reads in the dataset to ensure comparability between the two datasets (Validation cohort: 43 million RNA sequencing on-target reads, on average). MiOncoCircDB circRNA data were downloaded from the MiOncoCircDB webpage (https://mioncocirc.github.io/download/) and normalized using the same approach. For MiOncoCircDB sample classification, the available data were subdivided into four categories based on their respective tissues of origin: cancer_CNS, including all CNS malignancies; cancer_non_CNS, including all non-CNS malignancies; and the two healthy tissue categories, healthy_CNS and healthy_non_CNS.

4 Results

4.1 Development and application of the *circs* workflow

Circs is a RNA-Seq based circRNA quantification multi-pipeline approach similar to CirComPara[103] or Docker4Circ[118].

Circs utilizes three already published and established circRNA detection pipelines[53,99,119], find_circ, DCC and CIRCexplorer1, employing a subsequent three out of three confirmation vote between these pipelines, resulting in DCC-based quantifications of circRNAs found by all three pipelines.

4.1.1 Selection of appropriate circular RNA detection pipelines

The three pipelines were chosen based on multiple parameters and constrains: find_circ was the first published start-to-end pipeline. Here, find_circ was used to compare to old results (where only find_circ was used) and check the output of all other pipelines. find_circ also has a low ressource demand[100] and a relatively low false-positive rate, especially if the "40x40" filter is applied that can be used in the *circs* pipeline as a non-default option[102]. Find_circ, as other detection methods[88,108,109], provides a complementary database for circRNAs found with the algorithm[110]. This database was later used as a part of the circRNA candidate annotation to map identified circRNAs to circbase ids, directly adding one layer of information to the output of the pipeline. If needed, this can be used to filter for circRNAs that are only included in circbase. This feature was mainly used for validation purposes. DCC uses an even more resource-efficient mapper, STAR[47], with a high true positive rate[49] approaching 97% with a default minimum reads filter. Additionally, DCC has a high complementary score when coupled with find_circ[102]. The output of DCC can be used with FUCHS[112] and circtools[105] for back-splice junction PCR primer design and other downstream analysis of the identified circRNAs. CIRCexplorer1 has a high sensitivity used in combination with DCC and a high complementary score paired with find_circ.

These three pipelines can also be successfully applied to analyze single-read data, allowing the evaluation of additional datasets. CIRCexplorer1 is not able to detect *de novo* spliced circRNA species[102] and thus excludes some circRNAs from the output file generated by *circs*, for example ciRS-7[124]. Overall, the combination of three pipelines results in a more robust and sensitive analytical pipeline that is not dependent on the detection of many backsplice junctions in order to achieve an acceptable false-positive rate. Figure 8 offers a visual representation of the internal data flow of the created pipeline. As an additional function of *circs,* the "voting" step can be omitted and "unvoted" output of each pipeline can be used, which allowed for the identification of ciRS-7, for example[102]. Furthermore, DCC and find_circ are unable to detect non-canonical splice signals, thus precluding identification of another potentially interesting type of circRNAs[49]. This limitation is overcome by only using the output data of CIRCexplorer1 for downstream analysis.

4.1.2 *Circs* internal data flow

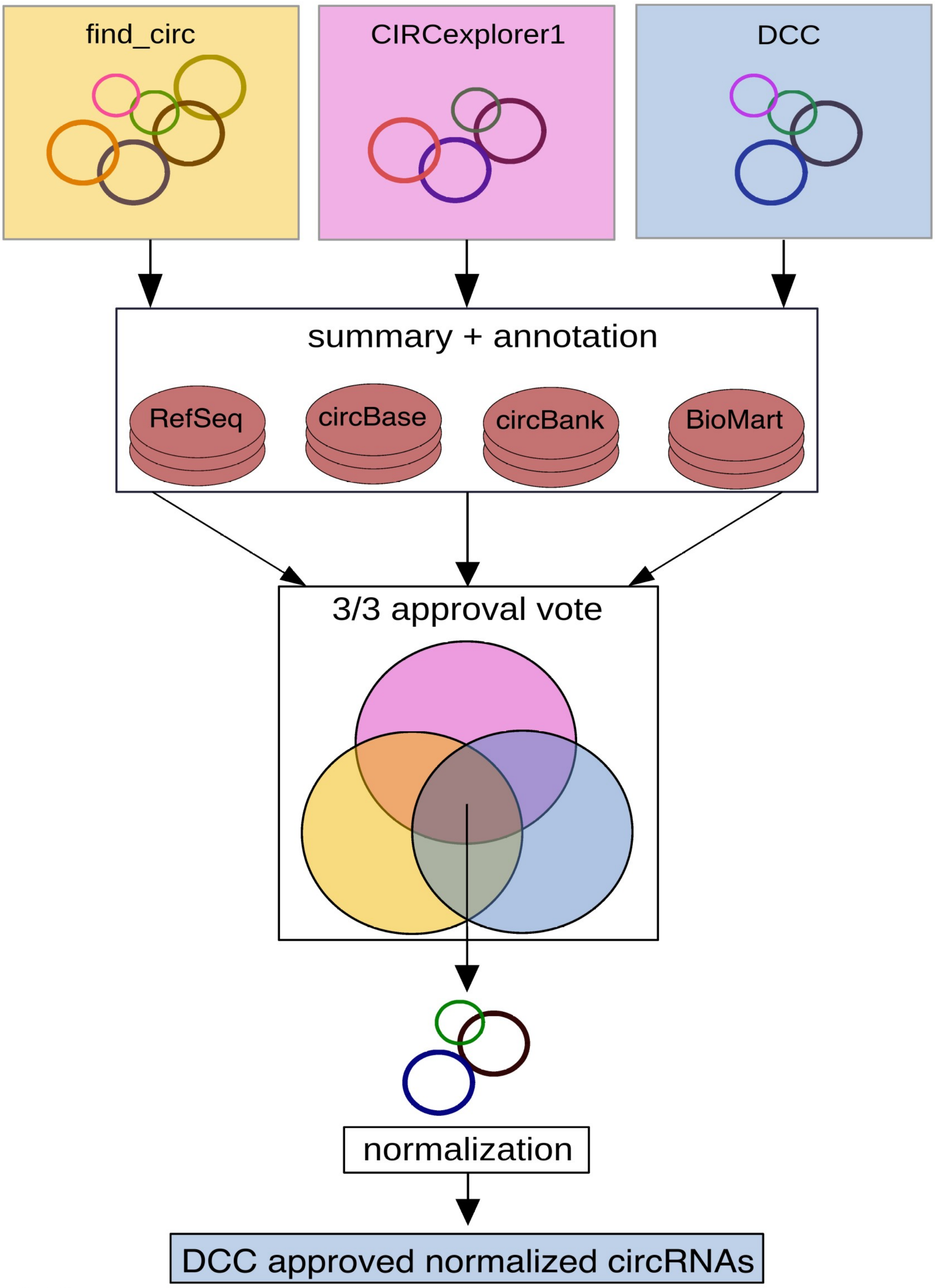


Figure 8: The general data flow of *circs* including the main steps pipeline run, summary and annotation, vote and normalization. First all three pipelines were run on the same input data and filtered according to the minimal reads threshold (two back splice junction reads). Next these results were subsequently summarized and annotated with information from circbase, circbank, RefSeq and BioMart. In a third step, the output circular RNA candidates of all three pipelines were overlapped and only putative circular RNAs that were found in all three filtered output files were accepted. In a fourth step the DCC quantifications of accepted (voted) coordinates were normalized to junction reads per million total RNA-Seq reads.

First, three circRNA detection pipelines were run on our RNA-Seq dataset. Two non-default options were chosen for the three pipelines: STAR was selected as the mapper for CIRCexplorer1, due to speed and accuracy[47]; and DCC's read-filter was set to two reads in at least one sample to improve interoperability.

Next, the resulting data from each pipeline were summarized, with circbase[110] ids, parental gene names and information as well as circbank[108] information if possible, added in the process.

In a third step called "voting", the output of the pipelines was compared and only circRNA species (genomic coordinates) that were detected by all three pipelines were approved. This step was crucial as it reduced the false positive rate, while preserving the sensitivity of each pipeline[100,102].

Normalization constitutes the fourth step in this workflow and improves the raw output of the circRNA quantification to make the results more comparable to other data analysis flows[50]. The raw back-splice junction read counts for each circRNA were divided by the total sum of RNA-Seq reads in the Fastq input files (in millions). This ensured comparability between datasets. *Circs* is able to process single-end and paired-end sequencing data, although paired end (pe) is the default data input for the pipeline and results in a higher degree of uniquely identified circRNAs in a given dataset (data not shown). The pipeline can be used with any genome if all annotation and reference files are available. So far, the pipeline has only been used with human reference genomes hg19 and hg38.

The *Circs* code (https://gitlab.com/daaaaande/circs) is written in Perl and based on automation scripts developed previously for a local server deployment: https://github.com/daaaaande/auto_find_circ, https://github.com/daaaaande/automate_DCC and https://github.com/daaaaande/circexplorer1_auto. These versions were only used on hg19 and are outdated. *Circs* was written and adapted for the local HPC environment called HILBERT at the Heinrich-Heine Universität Düsseldorf (HHU). The recently developed snakemake[125] based pipeline is available at https://github.com/daaaaande/circs_snake. This rewrite aims to improve access for other users and reproducible results, without dependency on the software packages included with HPC.

4.1.3 Execution of *circs*

Circs accepts .fastq files from single-end and paired-end sequencing. Both .fastq file types are trimmed and quality checked/filtered. Examples of RNASeq data quality checks are provided in Figure 2 and Figure 3. The file and sample annotations are provided as an infile.tsv file. This file can also be created with run_prep_guide.pl as an interactive run preparation tool. This tool was established in order to work within the local HPC rules and environment, therefore translation of this script is needed for use in other environments.

The primary pipeline runs can be started with pbs_array_execution.pl as a PBS array job. After completion of the PBS array job, the first step of this workflow is completed and is available for the next data analysis step. Subsequently, the script run_post_guide.pl concatenates output files of all samples for each pipeline and executes matrixmaker-V4.pl and matrixtwo_V4.pl consecutively for each included pipeline. Matrixmaker-V4.pl first summarizes the results of circRNA identification and creates a matrix including output for all samples from each pipeline. In the summarizing step, information about the included circRNAs is added in the form of genomic annotation of the circRNAs parental gene region. The second script, matrixtwo_V4.pl, adds further annotation information. Voting of the matrixtwo_V4.pl output file starts using auto_voting.R and data are subsequently normalized to junction reads per million using run_prep_guide.pl or a reads_per_sample.tsv file provided by the user. This file includes the total number of RNA-Seq reads for each analyzed sample and their sample names. The normalization is executed using norm_a_voted_circs_df.R, but any of the voted circRNA output files can be normalized separately, including CIRCexplorer1 and find_circ. A more detailed manual is available in the README.md file provided in *circs*.

The minimum reads threshold filtering was originally part of the output filtering step, but

HPC deployment has since been incorporated into the pipelines directly in order to decrease the output of unnecessary data.

As mentioned above, the vote is mainly used to minimize the false positive rate at the coordinate level. The exact circRNA candidate is defined by its unique genomic coordinates. The circRNA vote is based on overlapping genomic coordinates of all three pipelines and only accepts circRNA candidates that are detected by each pipeline. Once a set of coordinates is accepted by the vote, *circs* provides three normalized quantifications for the same set of coordinates from find_circ, CIRCexplorer1, and DCC. The user can choose which output file to use. In the own experience, the DCC output was the most sensitive when using approved circRNA coordinates and it found more back-splice junctions in the same data.

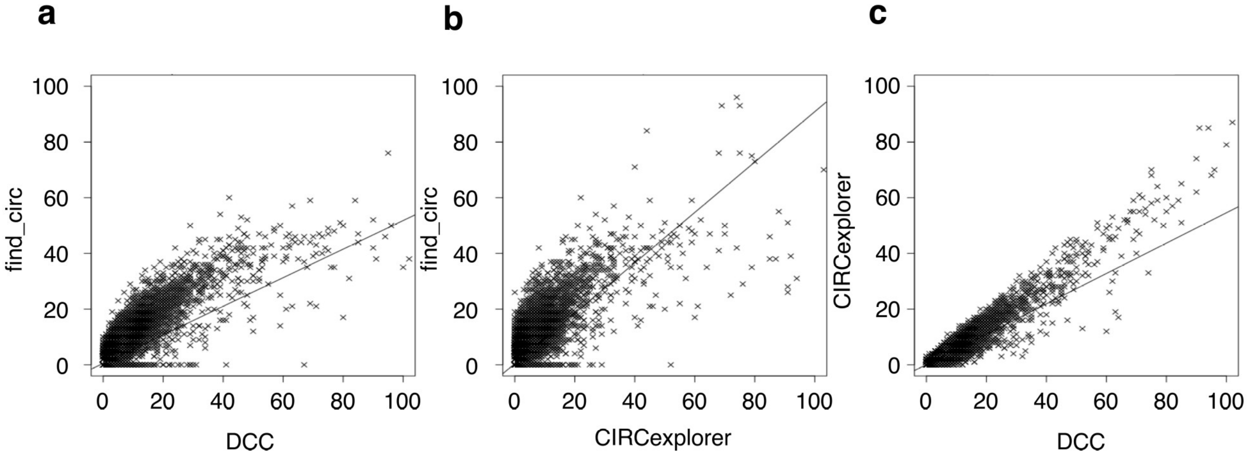


Figure 9: Correlation of raw circular RNA approved quantifications from the three datasets DCC, find_circ and CIRCexplorer1 correlated to each other. Linear trend line added. a) find_circ *versus* DCC. b) find_circ *versus* CIRCeplorer1. c)CIRCexplorer1 *versus* DCC. Each axis are raw circRNA junction quantifications.

This output of *circs* demonstrates the correlation of each of the datasets to each other (Figure 9). The find_circ pipeline utilizes a different RNA-Seq mapper (Bowtie2) than DCC and CIRCexplorer1. DCC is able to detect a greater number of approved circRNAs compared to CIRCexplorer1. The Pearson's correlation between find_circ and DCC is 0.77, between CIRCexplorer1 and find_circ 0.78 and between CIRCexplorer1 and DCC (both using STAR for RNA-Seq mapping) 0.94.

Circs output format is a .csv file with the normalized circRNA reads for all samples included in a run, with annotations added to each circRNA, such as RefseqID, host gene name, strand information, circbase id, cancer hallmark association of the parental gene, biological description of the parental gene, and, if desired, coding potential of the circRNA according to circBank and mm9 (mouse) homolog circRNA.

4.1.4 Circs_snake

To leverage the functions of snakemake, a snakemake-based version of circs was created: circs_snake.

The snakemake-based circs_snake workflow is available at https://github.com/daaaaande/circs_snake and part of the snakemake workflow catalog.

Circs does not come with software versions of all packages used, but if it is setup to use the same software that circs_snake comes with, the output is identical. The circs_snake workflow is intended to be published later.

Circs_snake introduces several changes compared to circs to ease the use and improve the reproducibility of the pipeline based on snakemake features:

Most required packages are available via conda (www.anaconda.org) and thus can be managed with conda. This greatly simplifies the setup process and makes the installation and management of the used tools easy. The conda packages required for each step are listed in .yaml files that can be found in the envs/ directory of circs_snake. These .yaml files and their included packages can be handled by snakemake, installing all tools during the first run of the pipeline. Two of the three circRNA detection pipelines are the exception here, as the required software is not available through conda. These tools, DCC and find_circ, still need to be installed in other ways. Needed aligner-specific reference genome indices also need to be created before a first run, or provided in the config.yaml file.

Additionally, snakemake supports “dry-runs”. These “dry-runs” are a theoretical run of the user’s setup where no rules are executed, but the exact commands that would be executed outside a dry run are listed with input files, output files, software environment that will be used and the reason why a specific rule is executed. This is especially helpful when preparing a pipeline run and to catch any abnormalities that might occur.

Due to the structure of circs_snake, files that are created by any rule can be defined as final output files, making the whole workflow more versatile if only certain output files are needed. Rules are only executed if their output files are needed by other rules or are stated as the desired final output. The same feature is used in the first few steps of the pipeline: the three utilized circRNA detection tools require uncompressed .fastq files as input, and circs_snake is able to detect wether the input files need to be decompressed (from .fastq.gz files) or not and act accordingly.

Circs_snake is also able to conveniently switch between environments with job scheduling software and other, usually smaller environments without. Since its first deployment circs_snake is used on the local HPC with PBS Pro as a managing layer between many users and finite computing resources. If the scheduling-specific snakemake parameters are not given during the pipeline execution, circs_snake will execute the rules without one. In the current HPC based setup, each rule is executed as a single PBS Pro job. For each job, separate logfiles are created by PBS Pro for output and error messages, giving the user additional context for each job.

All genome index, circRNA annotation files and deployment-specific directories need to be customized in the config.yaml file. From this file circs_snake retrieves files, folder locations and parameters needed for the pipeline execution. This config.yaml file can be cloned and changed for each genome of interest or different setups. This enables easy and fast switches between runs of the pipeline for different organisms, if needed.

The file cluster.json includes the resources for each job that will be requested. This file includes one general setup for minor tasks, but custom setups for each resource intensive part.

Snakemake organizes the rules that need to be executed into a directed acyclic graph (DAG). This structure results in an independent execution of several parts of the pipeline. This is especially helpful if some samples in one dataset are corrupted or differ in size compared to the other samples of the dataset. If one rule is finished, the used resources are freed directly afterwards and the resulting files are ready for the next rule. In the non-snakemake based circs there is only one linear path of command execution until the final output is generated or one error is encountered, with one set of resources allocated to one sample for all steps combined. For the mapping step 12 CPU cores are used by default, resulting in 11 unused CPU cores during most other steps. Conversely, the snakemake approach of defining resources for each rule results in less total resources used and thus a faster total execution time. Each rule that needs only one CPU core will occupy only one CPU core. Additionally, If one rule fails or the complete snakemake run is interrupted, the run can be resumed later without additional checks – snakemake notifies the user if rules were executed unsuccessful or output files are incomplete, with the option to repeat incomplete rules the workflow can be continued if desired.

To illustrate all steps taken in a theoretical circs_snake run with one sample, the DAG figure generated by snakemake including all to be executed rules is shown below in Figure 10.

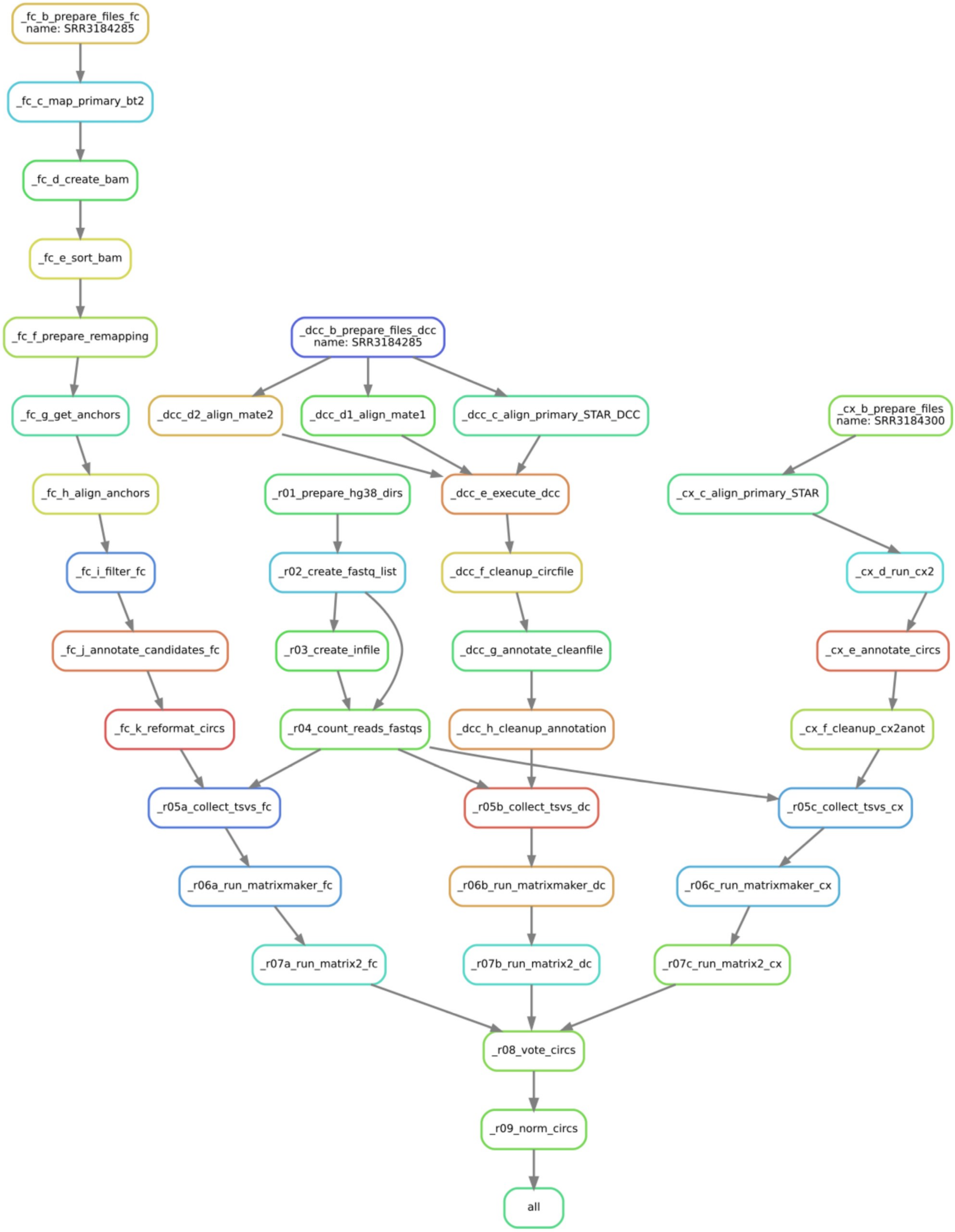


Figure 10: Circs_snake DAG*.* General rules are denominated by a number, workflow specific rules by dcc ,fc and cx for DCC, find_circ, CIRCexplorer respectively.

This DAG can be split into four different parts:

The rules starting with _r0 are the pipeline independent steps: preparation, read counting for later normalization and collection of output of the three circRNA detection pipelines. Here output folders are created, reads for each sample are counted and later used for circRNA count normalization and the 3/3 vote is executed. General rules are numbered while pipeline-specific rules are denominated with letters. Circs_snake is also split into four separate snakemake files as represented by the rule names. One for each circRNA detection tool and a fourth one for all general rules.

Rules starting with _fc are the find_circ specific rules. After a count file including all detected back-splice junctions for each sample is created, results are converted into a count matrix file including results from all samples.

The DCC pipeline is executed by rules starting with _dcc. Here each of the two read files is mapped separately to the reference genome in addition to the combined mapping. The results of these three mapping steps are then collected in step _dcc_e.

CIRCexplorer steps are denominated with the prefix _cx. This part of the pipeline is the least resource intensive as it includes only one mapping step, less steps overall and the more resource efficient mapper of the two used, STAR.

The overall DAG structure shows a difference between find_circ and the two other tools, namely find_circ includes more steps than DCC and CIRCexplorer. Another notable difference is the number of cores used by these tools outside the mapping step: find_circ uses one CPU core while the other two tools can use multiple.

The rule “all“ is the final rule of circs_snake. It takes three voted and normalized circRNA tables as an input and does not create any output. Circs_snake is not a complete re-write: many scripts have been taken from circs. For example, the voting and normalization of the voted circRNAs are the same script in both pipelines that gets executed. Additionally, scripts that reformat each pipelines output and create the circRNA count tables are also not specially created for circs_snake.

4.1.5 Circs_snake execution

To give an insight into the circs_snake execution, a typical execution command is shown below, with each part explained separately below.

```
snakemake -p --cluster-config cluster.json --cluster "qsub -A {cluster.account} -q
{cluster.queue} -l walltime={cluster.time}-l
select={cluster.nodes}:ncpus={cluster.ncpus}:mem={cluster.mem}:arch={cluster.arch}" -j
100 --latency-wait 90 --use-conda --max-status-checks-per-second 1 --keep-going
```

snakemake -p : is the snakemake command itself. The -p results in snakemake showing all commands that are executed. This part is the only mandatory part of the snakemake execution. All other options are optional and can be combined as needed. The -p argument is also not needed, but is highly recommended to be used.

--cluster-config cluster.json : points snakemake to the file listing cluster-specific resource allocations for each rule.

--cluster "qsub -A {cluster.account} -q {cluster.queue} -l walltime={cluster.time} : shows how the interaction with the job-scheduling software, in this case PBS Pro, happens. The curly brackets will be filled in with numbers and strings from the aforementioned cluster.json file, differing between rules.

-j 100 : This parameter limits the number of concurrent job submissions to PBS. A higher number can result in PBS instability issues.

--latency-wait 90 : When jobs are executed, snakemake checks the outputfiles. On distributed file systems like the ones used for HPC, big files are not instantly synchronized between nodes but need time. This lets snakemake wait for 90 seconds before assuming a file is missing.

--use-conda : This enables snakemake to install missing packages for each environment using the conda package manager.

--max-status-checks-per-second 1: Snakemake also checks if jobs are executed successfully, are still waiting in the jobs queue to be executed or are already finished without errors. This parameter is used to limit the amount of job status checks to one check per second.

--keep-going : If a single job failed, snakemake can either stop the pipeline completely (not submitting any more jobs) or as shown here, continue with its execution until all other executable (meaning that all input files for these rules are created or still can be created) rules are done.

4.2 Discovery cohort

The first dataset analyzed with *circs* was taken from Forget et al. 2018[33]. The initial cohort comprised RNA-Seq data from 38 primary MB samples, for most of which additional DNA methylation, proteomics and phosphoproteomics data were available. Further characteristics of this dataset are provided below.

4.2.1 Detection of circular RNAs in the discovery cohort

The trimmed RNA-Seq reads of the discovery cohort were taken as input for circRNA analysis with *circs*. The initial numbers of uniquely identified circRNAs by each of the three pipelines is illustrated in Figure 11.

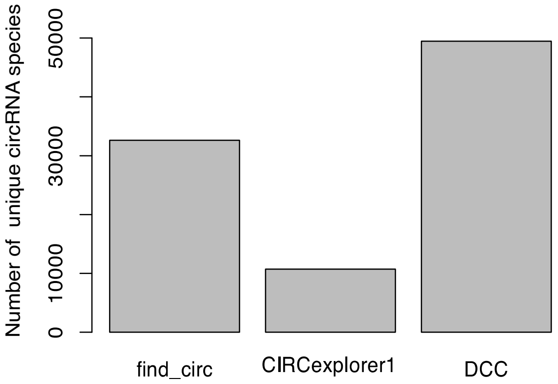


Figure 11: Amount of unique circular RNAs identified from each pipeline in the discovery cohort. DCC detected the most circular RNAs, find_circ an intermediate amount, while CIRCexplorer1 detected the least.

This plot illustrates the absolute number of circRNA species in the discovery cohort data uniquely identified by each of the three pipelines. CIRCexplorer1, DCC and find_circ detected a total of 10724, 49470 and 32642 unique circRNAs, respectively. The absolute numbers of detected circRNAs depended on the applied filter settings. Specifically, a unique circRNA had to be detected twice in at least one sample to be listed. These results were then used for the “voting” step illustrated below in Figure 12.

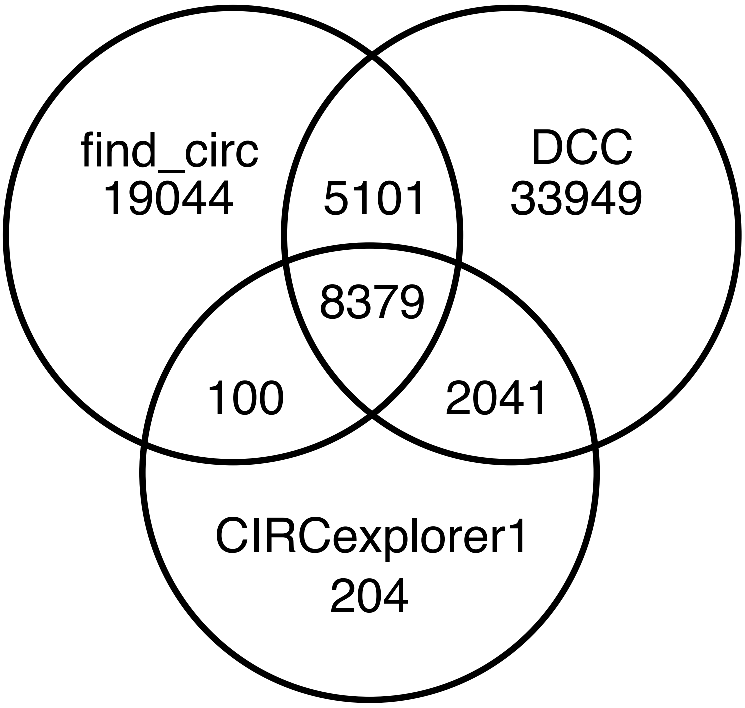


Figure 12: The vote result of the discovery cohort. The vote including all three algorithms agreed on 8379 circular RNA species. The DCC quantifications of these circular RNA candidates were normalized and then used for all subsequent analysis.

The Venn diagram shows that the voting step resulted in 8379 accepted unique circRNAs. The overlaps between pairs of the three datasets were asymmetric. Most of DCCs coordinates (68.62%) were non-overlapping, as expected, since DCC detected more circRNA candidates than the other pipelines. By contrast, most CIRCexplorer1 coordinates overlapped with at least one of the other algorithms (only 1.9% are non-overlapping). CIRCexplorer1 identified a smaller absolute number of circRNA candidates compared to DCC and find_circ. The results obtained by find_circ were similar to those by DCC: 58.34% of putative circRNAs detected by find_circ are in no other dataset and thus do not overlap. A total of 7242 putative circRNAs were identified by only two of the three pipelines. The number of accepted circRNAs would have been almost twice as high if *circs* had accepted putative circRNAs detected by two out of three algorithms. This type of voting is employed in CircComPara, where only two out of several algorithms must agree. Thus, *circs* constitutes a circRNA detection pipeline with relatively strict filtering criteria. The quantifications from DCC of the voted coordinates were subsequently normalized. The resulting data was called the discovery cohort.

4.2.2 Medulloblastoma group classification utilizing circular RNA data in the discovery cohort

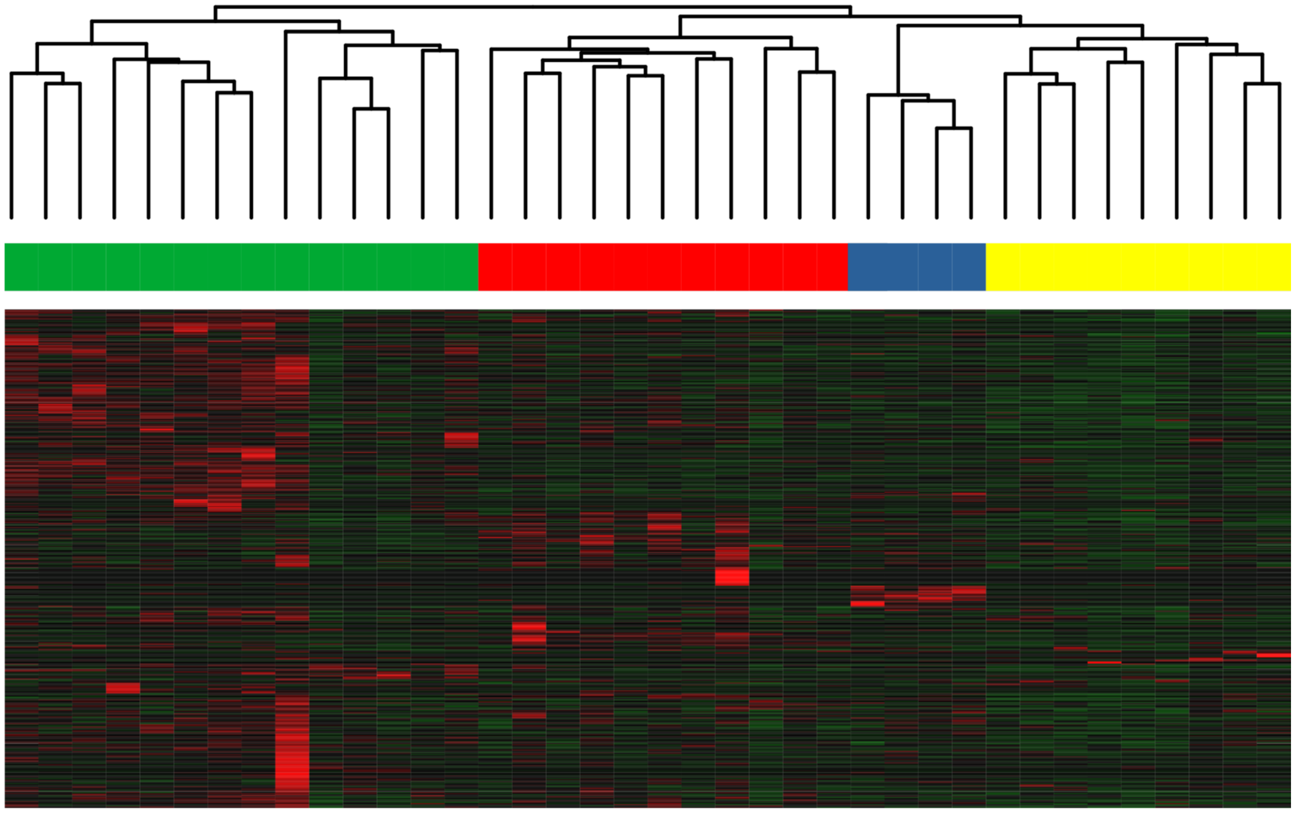
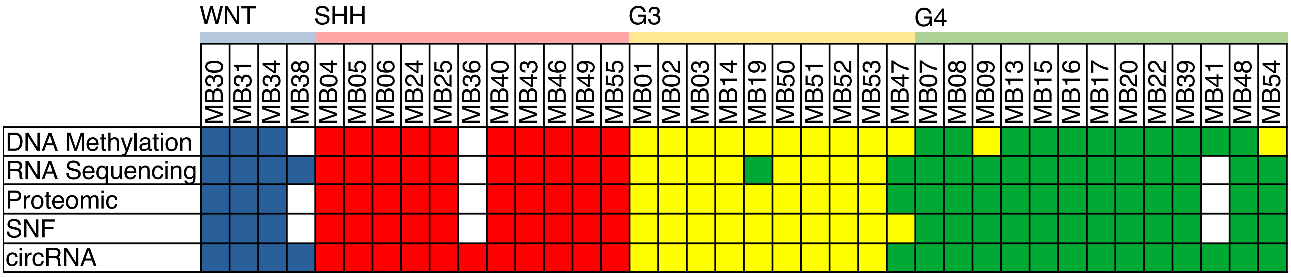


Figure 13: Heatmap of top 500 differentially expressed circular RNAs across all 38 samples in the discovery cohort. Top colors and dendrogram according to circs' clustering (average Pearson's dissimilarity). Blue= WNT medulloblastoma, red = SHH medulloblastoma, yellow = Group 3 medulloblastoma, green = Group 4 medulloblastoma.

To gain insight into the discovery dataset and the group-wise clustering, a heatmap was created based on the expression levels of the top 500 differentially expressed circRNAs across the dataset. The result is depicted in Figure 13. The dendogram above the heatmap revealed an initial separation between Group 4 MB and a group of WNT, SHH and Group 3 MBs. Compared to other omics studies, this was an uncommon initial separation of MB data[14,17,24,29(p5)]. The second degree of separation was observed between SHH and Group 3 and WNT MBs. The final MB group wise separation was determined between WNT MB and Group 3 MB. Furthermore, several subclusters could be observed in SHH, Group 3 and Group 4 MBs. One Group 4 MB sample (MB07) demonstrated aberrantly high circRNA expression for a large set of circRNAs.

The depicted hierarchical clustering result of the top 500 differentially expressed circRNAs was compared with the previously published sample classification according to several methods, including SNF[126](Figure 13).



	WNT				SHH											G3										G4												
	MB30	MB31	MB34	MB38	MB04	MB05	MB06	MB24	MB25	MB36	MB40	MB43	MB46	MB49	MB55	MB01	MB02	MB03	MB14	MB19	MB50	MB51	MB52	MB53	MB47	MB07	MB08	MB09	MB13	MB15	MB16	MB17	MB20	MB22	MB39	MB41	MB48	MB54
DNA Methylation	🟦	🟦	🟦	⬜	🟥	🟥	🟥	🟥	🟥	⬜	🟥	🟥	🟥	🟥	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨
RNA Sequencing	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	⬜	🟥	🟥	🟥	🟥	🟥	🟨	🟨	🟨	🟨	🟩	🟨	🟨	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩
Proteomic	🟦	🟦	🟦	⬜	🟥	🟥	🟥	🟥	🟥	⬜	🟥	🟥	🟥	🟥	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩
SNF	🟦	🟦	🟦	⬜	🟥	🟥	🟥	🟥	🟥	⬜	🟥	🟥	🟥	🟥	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩	🟩
circRNA	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Figure 14: Comparison of group classification from Forget et al. 2018 and top 500 differentially expressed circular RNAs (Pearson's dissimilarity) in the discovery cohort. Similarity network fusion (SNF) is a method using multiple input data layers, resulting in one consensus classification. When a sample lacked any omics data layer, SNF could not be performed. Blue = WNT medulloblastoma, red = SHH medulloblastoma, yellow = Group 3 medulloblastoma, green = Group 4 medulloblastoma, White boxes = no data available.

Figure 14 shows a broad agreement between most clustering approaches using distinct input data, but with some key differences: MB38, MB36 and MB41 did not have an SNF group annotation but are included in the circRNA analysis. This is due to several missing omics layers for these samples, which are necessary to determine the SNF-based group. MB38, for example, showed insufficient RNA-Seq data quality and had to be excluded from the RNA-Seq data analysis. However, this was not the case for circRNA analysis, for which the data quality was sufficient. MB36 was diagnosed as SHH MB by a pathologist in the standard clinical setting and, due to sufficient RNA-Seq data quality for circRNA analysis, was also included in this dataset. MB41 showed low cellularity in the sample and the sample quantity did not suffice for most analyses. However, the quantity and quality of RNA-Seq data was still sufficient for circRNA analysis. These cases illustrate the two major advantages of circRNA-based clustering of these samples: More samples could be included since other omics data layers were not needed, with the exception of RNA-Seq data, and lower RNA quality was acceptable. Notably, MB36 was not classified into one of the four MB groups by any other omics method, with the exception of circRNA data. To further understand the resulting sample clusters in the discovery cohort, the detected circRNA signal for each sample was sorted into the previously described MB groups based on SNF, and again based on the circRNA MB groups defined here (Figure 15).

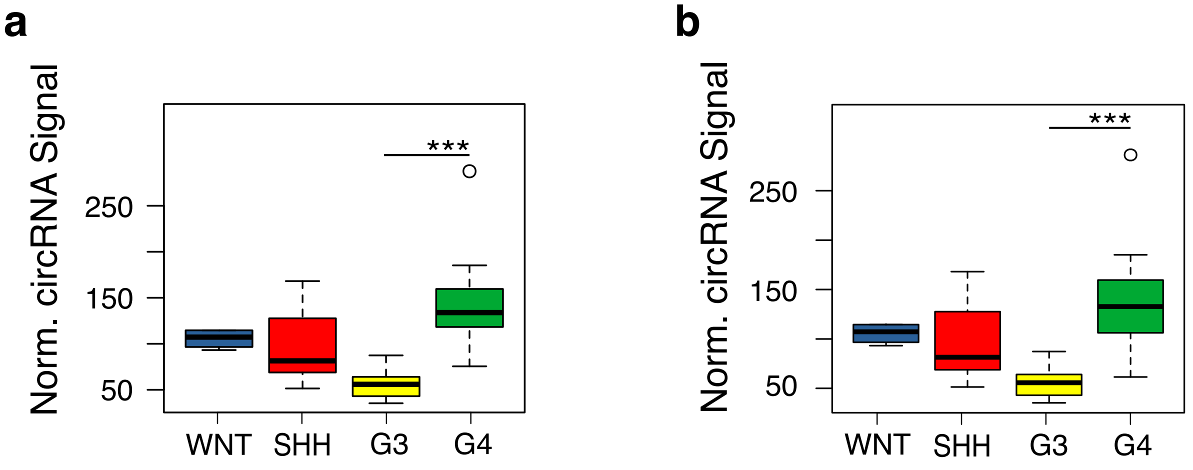


Figure 15: Circular RNA signal across medulloblastoma groups according to two differing MB group classifications. a) Circular RNA signal according in SNF MB groups. b) Circular RNA signal in MB groups according to circular RNA data. ***=Tukey HSD p-value adjusted <0.001. Normalized circRNA signal to back-splice reads per million.

These boxplots show a highly significant difference in the total circRNA signal between Group 4 and Group 3 MB (Tukey HSD p-values adjusted = 4.1x10^{-4} for SNF-based MB groups and 2.9x10^{-4} for circRNA-based MB groups). The outlier in Group 4 in each of these plots was MB07, which was determined to be a Group 4 MB according to SNF and circRNA data analysis. Significant differences in the mean circRNA signal were observed between Group 3 and Group 4. Notably, Group 3 MB had particularly low normalized circRNA signal and that of Group 4 MB was particularly high.

To further determine MB-group-specific circRNA expression patterns, a second cohort, called the validation cohort, was analyzed using the same approach as was used for the discovery cohort.

4.3 Validation cohort

The validation cohort included previously unpublished RNA-Seq data from 35 MB samples. At the time of the circRNA analysis, only DNA methylation and RNA-Seq-data-based MB groups were known, so no SNF result could be obtained for these samples. After the voting step in *circs*, 8460 circRNAs were accepted, and the RNA-Seq and DNA methylation MB groups were identical in this dataset for each sample. The circRNA quantifications of approved DCC coordinates were then normalized to back-splice junction reads per million total RNA-Seq reads, as previously described for the discovery cohort.

4.3.1 Medulloblastoma group classification utilizing circular RNA data in the validation cohort

The top 500 differentially expressed circRNAs in the validation cohort were clustered according to Pearson's dissimilarity and a heatmap was generated, as described above for the discovery cohort. The clustering result is shown below in Figure 16.

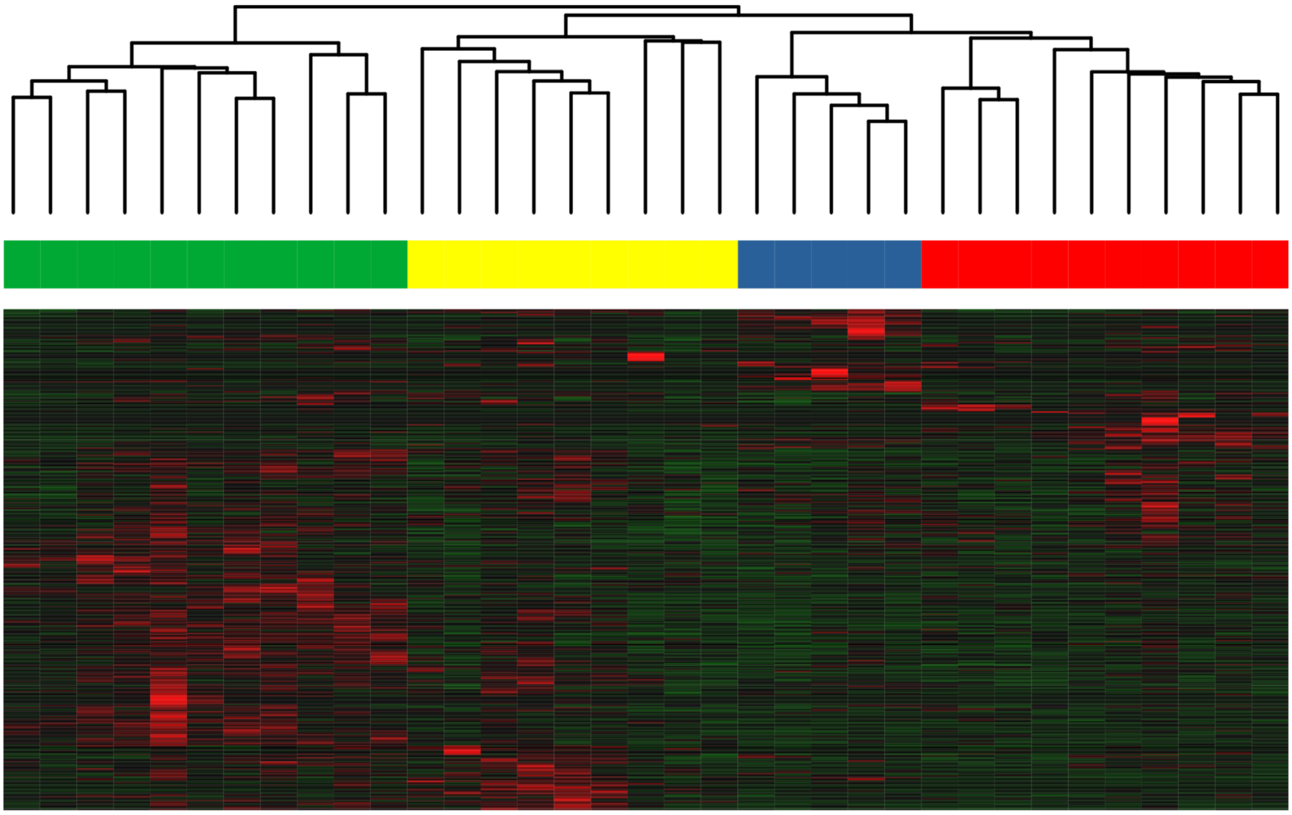


Figure 16: Heatmap of top 500 differentially expressed circular RNAs across all 35 samples in the validation cohort. Top colors and dendrogram according to circular RNA group classification (average Pearson's dissimilarity). Group colors according to discovery cohort legend (blue = WNT MB, red = SHH MB, yellow = Group 3 MB, green = Group 4 MB).

Similar to the discovery cohort clustering results, Group 4 was first segregated from all other MB groups according to the hierarchical clustering dendrogram of the circRNA data shown in Figure 16. Notably, the discrimination of the other MB groups was different compared to the discovery dataset: Group 3 MB separated from SHH MB and WNT MB first, while WNT MB and SHH MB segregated in the last step when distinguishing MB groups. Next, circRNA signal strength was compared between the four MB groups in the validation cohort (Figure 17).

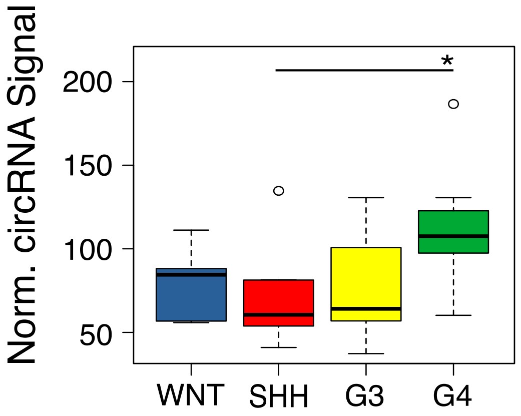


Figure 17: Circular RNA signal across medulloblastoma groups in validation cohort. Here the difference between Group 4 and SHH MB was significant.*= Tukey HSD p-value adjusted <0.05.

The samples of the validation cohort were annotated according to circRNA, DNA methylation and RNA-Seq MB group classification. MB_V9 was the top outlier in Group 4, with the highest circRNA signal across the whole dataset. The only significant difference according to MB subgroups in circRNA signal was seen for Group 4 and SHH MB (p-value adjusted Tukey HSD = 0.0208), while in the discovery cohort, the only significant difference was observed for Group 3 MB and Group 4 MB. Notably, in both cohorts, Group 4 MB had the highest circRNA signal.

4.4 Cross-dataset circular RNA signal evaluation

To further compare the circRNA signal in MB and normal controls, a cohort of 12 healthy fetal brain tissues was analyzed as reported above for the discovery and validation cohorts. The resulting comparison of the normalized circRNA signal across these three datasets (healthy brain tissue, discovery and validation) is depicted below in Figure 18.

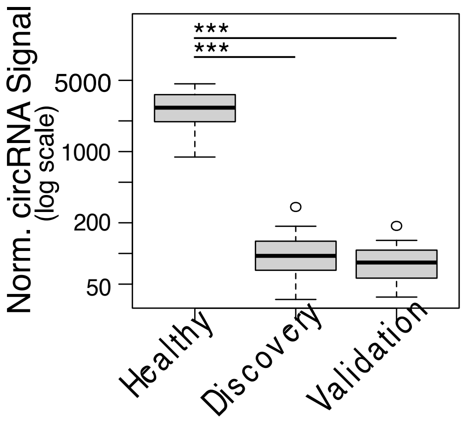


Figure 18: Circular RNA signal across three datasets. ***=p-value adjusted Tukey HSD <0.001. Healthy = ENCODE healthy fetal brain tissue samples.

The difference between the healthy brain, discovery and validation cohorts was highly significant (Tukey HSD p-value adjusted discovery *versus* healthy_brain < 2.62×10^{-14}, validation *versus* healthy_brain < 2.62×10^{-14}). Note the log-scale of the boxplot. The circRNA signal boxes of the discovery and validation cohorts showed only a minimal circRNA signal difference in comparison.

To further assess the similarity of the discovery cohort and validation cohort data, all included samples were correlated to each other. The resulting Pearson correlation matrix is represented in the corrplot below (Figure 19).

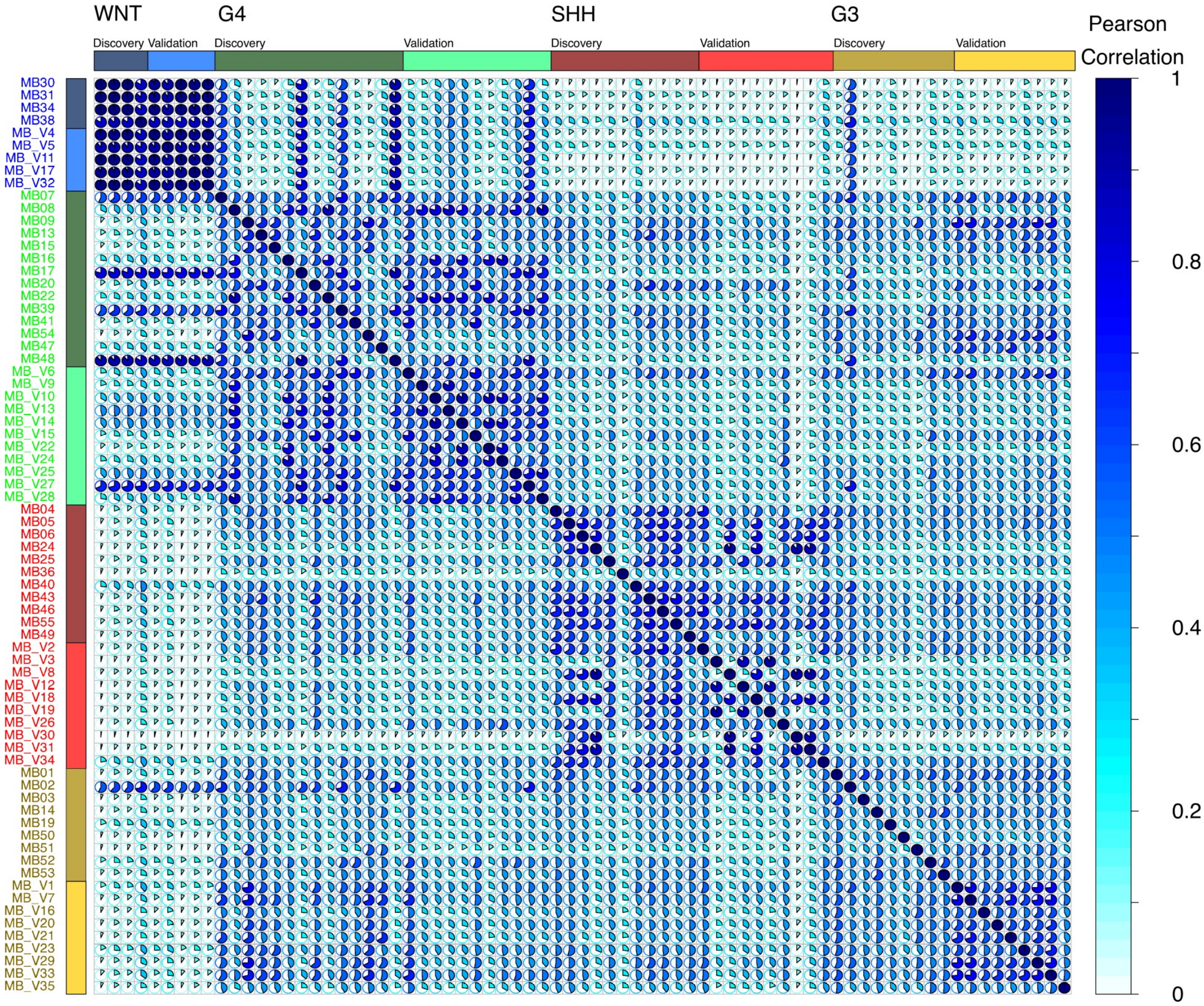


Figure 19: Corrplot of discovery and validation circular RNA data. Pie color and filling represents Pearson correlation. Higher correlation is represented by a darker color and a more filled circle. Sample order according to circular RNA MB groups and cohort.

This analysis demonstrated a strong correlation between WNT MBs throughout both cohorts, and the lowest correlation between any WNT MBs in this combined cohort was 0.824. Additionally, the four MB groups formed visible clusters in the combined dataset, with WNT MB clearly showing the highest degree of correlation of individual samples. The diagonal line in Figure 19 shows the correlation of each sample with itself and corresponded to the expected correlation coefficient of 1 in all cases. Samples that had a relatively high correlation with WNT samples were observed in each MB group. Samples MB02, MB48 and MB_V27 had a particularly high correlation coefficient with WNT MB subgroup samples, despite the fact that these samples were ultimately classified as Group

3 MB or Group 4 MB samples. After comparing the overall similarity of samples across these two cohorts, the data were investigated for potential circRNA biomarker candidates across both datasets, as well as in the discovery and validation cohorts separately.

4.5 Medulloblastoma pan-cohort group circular RNA biomarker definition

To detect biomarkers across the discovery and validation cohorts, all circRNA data were divided into their respective circRNA-defined MB groups, as previously described. An ANOVA was carried out for all circRNAs present in the dataset, followed by a Tukey HSD as a post-hoc test. The circRNAs that showed a significant difference (Tukey HSD p-value adjusted <0.05) in all comparisons between one group of interest and the three remaining groups in the discovery cohort were selected for evaluation in the validation cohort. Biomarker identification for each subgroup was performed using the same approach for the validation cohort. To reveal the amount of circRNAs with significant expression differences according to MB subgroups across both datasets, the two biomarker lists were compared (Figure 20). The resulting overlappeing circRNAs are also listed in supplementary Tables 11-13.

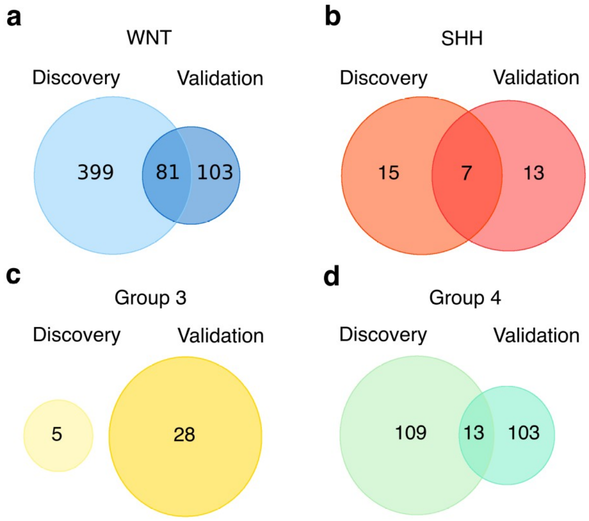


Figure 20: Group-specific biomarker identification in the discovery and validation medulloblastoma (MB) cohort. Significant circRNA expression differences according to MB group as determined by ANOVA followed by a Tukey HSD as a post-hoc test. a,b,c,d depict group-specific biomarker candidates in discovery and validation WNT, SHH, Group 3 and Group 4, respectively.

Initially, a strikingly different overlapping of potential biomarkers was observed: Most group-specific circRNA biomarkers were detected for the WNT MB group, while not a single overlapping circRNA-based biomarker could be identified for Group 3 MB. Since most group-specific circRNA biomarkers were determined for WNT MB, most overlapping circRNAs across both datasets also belonged to this group. Next, the mean expression levels according to MB group were compared to further investigate the biomarker potential of overlapping circRNAs candidates. This step aimed to exclude candidates with highly significant expression chances that were not biologically relevant, as some of these candidates had very low expression levels in all datasets. Only the candidates with relatively high expression are reliably detected, since circRNAs with low expression levels are more likely to be destroyed during sample preparation, simply do not reach the detection limit of the specific methodological platforms or suffer from a comparably high amount of linear mRNA noise.

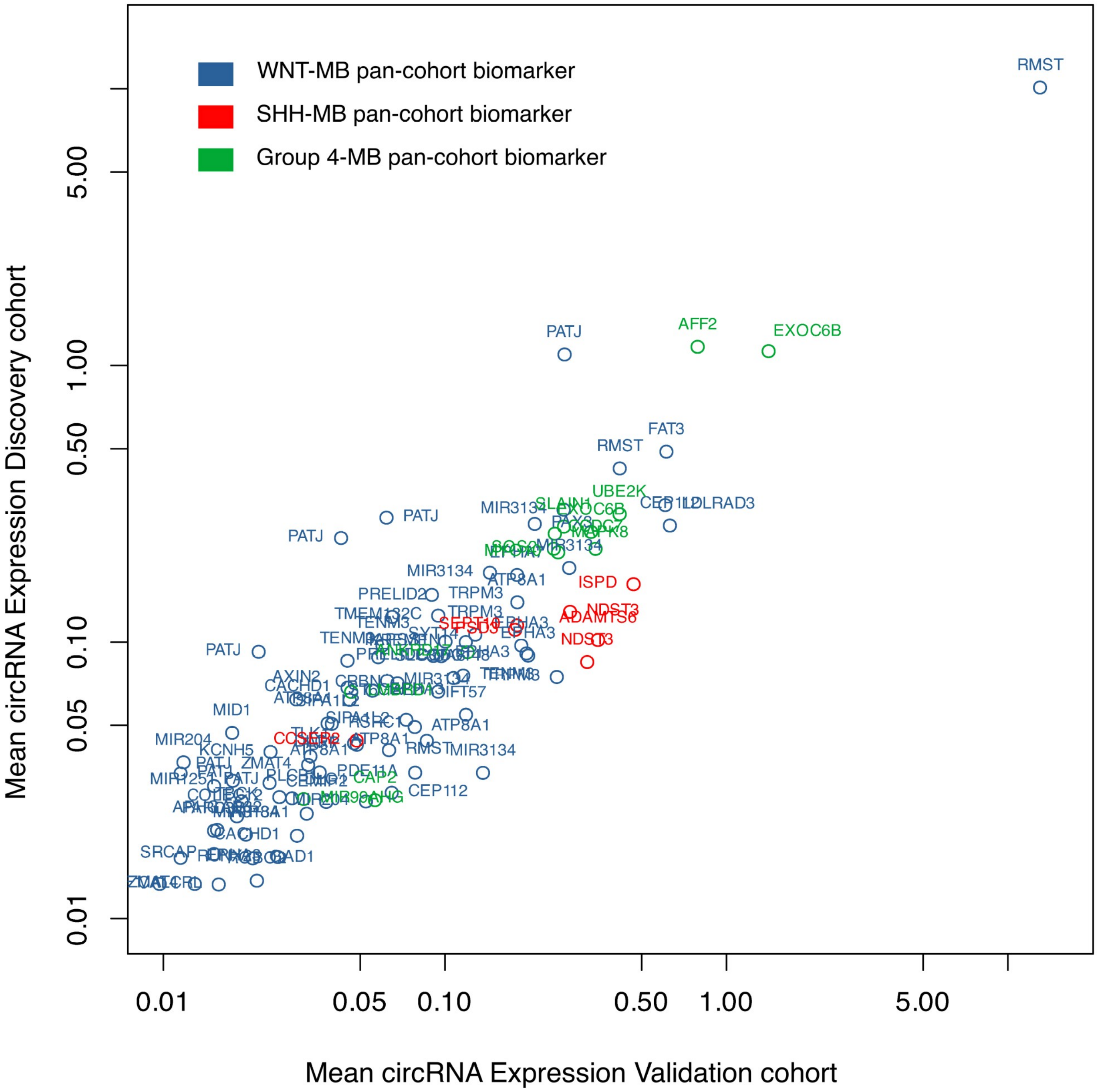


Figure 21: Mean in-group expression of circular RNA biomarkers in discovery cohort versus validation cohort. Logarithmic scale of normalized circRNA junction reads per million. Names above or next to the data points show the parental gene of each circRNA candidate (if the circRNA consits of exons from geneA, the label is geneA).

For three MB groups, pan-cohort circRNA biomarkers with the highest normalized mean expression were selected. However, this selection was not possible for Group 3, since no overlapping biomarkers could be identified when both datasets were compared. Multiple circRNA transcripts that originated from the same parental gene, including different exons, were identified as significant biomarkers for some genes (e.g. *RMST* and *PATJ*). The outlier in the top right corner of Figure 21 was circ*RMST*, consisting of exons 6-12 with the backsplice junction between exon 12 and exon 6. The other isoforms of circ*RMST* were significant as well, but all showed a lower mean expression levels compared to circ*RMST*$^{6-12}$. Due to the aberrant overexpression of this isoform of circ*RMST* in both datasets, circ*RMST*$^{6-12}$ (further called circ*RMST*) was selected for further investigation.

4.5.1 Circ*RMST* as WNT medulloblastoma biomarker

To investigate the MB-group-specific expression patterns of this circRNA candidate, its expression was first investigated in all samples of the discovery cohort.

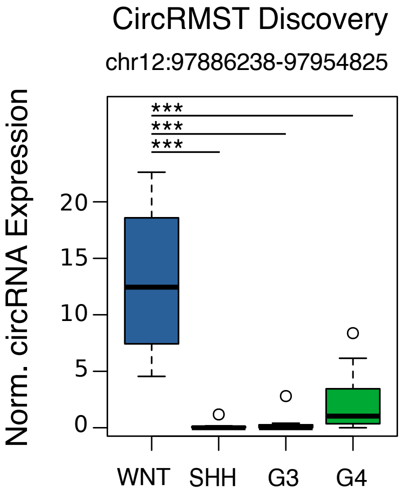


Figure 22: Boxplot of normalized circ*RMST* expression in the discovery cohort. Samples are ordered according to circular RNA medulloblastoma groups, genomic coordinates based on hg19. ***=Tukey HSD p-value adjusted<0.001.

Figure 22 shows aberrant circRMST expression in WNT MB in the discovery cohort, rendering it an interesting candidate for further investigation. Outliers with high circ*RMST* were observed in a small subset of non-WNT MB. To further confirm the biomarker potential of circ*RMST*, circ*RMST* expression patterns were determined according to MB groups in the validation dataset using the same approach.

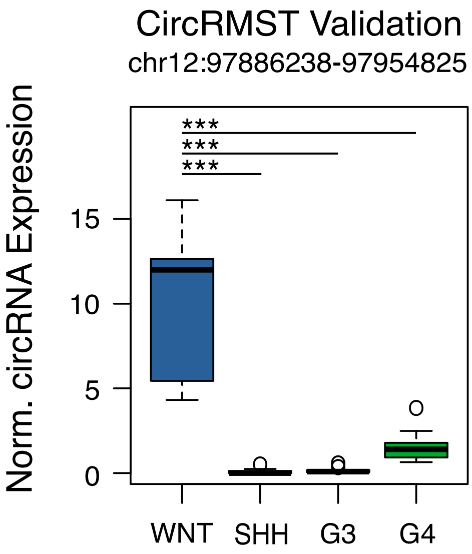


Figure 24: Boxplot of normalized circ*RMST* expression in the validation cohort. Samples are ordered according to circular RNA medulloblastoma groups, genomic coordinates based on hg19. ***=Tukey HSD p-value adjusted<0.001.

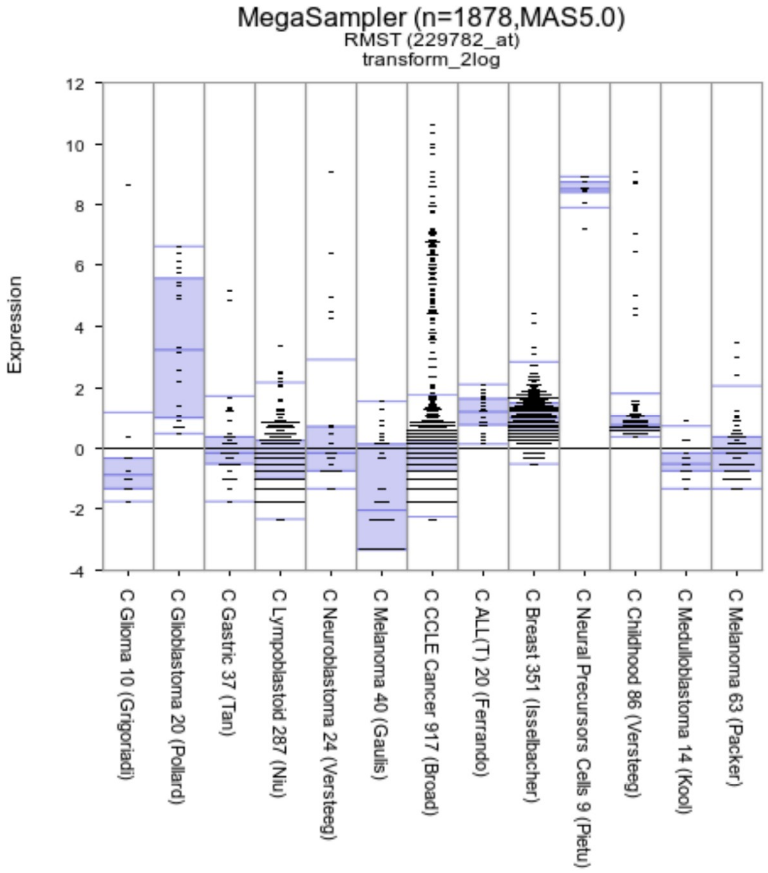


Figure 23: Megasampler of *RMST* RNA expression in several cancer data sets. Taken from https://hgserver1.amc.nl/cgi-bin/r2/main.cgi. CCLE=Cancer Cell Line Enzyclopedia, normalized expression values. A Megasampler is a boxplot comparing expression of one feature between multiple datasets.

The median expression level of *RMST* was lower in WNT MBs compared to WNT samples in the discovery cohort (mean circ*RMST* expression of 13.013 circRNA junctions/10^6 reads in the discovery cohort versus 10.101 circRNA junctions/10^6 reads in the validation cohort). However, the overall expression patterns remained highly comparable between the discovery and validation cohorts: Aberrant overexpression of circ*RMST* was detected in the WNT MB group compared to all other groups. In line with the discovery cohort, Group 4 displayed higher circ*RMST* expression compared to SHH MB and Group 3 MB. Furthermore, a small subset of SHH MB and Group 4 MB samples had relatively low expression but detectable expression of this candidate as outliers. Overall, circ*RMST* expression was strikingly similar in both datasets, underlining the biomarker potential of this circRNA for WNT MBs.

The parental gene of this circRNA, *Rhabdomysarcoma associated transcript 2* (*RMST)*, has been the subject of a number of studies. This non-protein coding gene is located on chromosome band 12q21. Its exact hg19-based genomic coordinates are chr12:97856554-97958793 on the plus-strand and the genomic sequence includes 14 exons. *RMST* has previously been associated with the development of the human brain and cancerogenesis [59,127–130]. In addition the circular RNA form was shown to be the dominant transcript compared to the linear form of this long non-coding RNA[131]. It was just recently demonstrated that the circular RNA reported here is expressed in MB and ependymomas[104]. Furthermore, circ*RMST* was previously also detected in the brain[59]. The linear form of the long non-coding RNA *RMST* has been found to regulated neurogenesis by interacting with *SOX2*[127]. Another publication showed a direct interaction between *RMST* (here a “trans-spliced” form, “ts-*RMST*”) and WNT[132], postulating a direct impact of ts*RMST* on non-canonical WNT-signaling. Our biomarker candidate circ*RMST* was detected in multiple studies according to circbase. To investigate the expression patterns of circ*RMST* in many additional datasets including different cancer entities, *RMST* expression was determined in databases that cannot distinguish between the linear and the circular isoform. Since the circular isoform of *RMST* constitutes the predominant isoform, this approach already provides an indirect indication of circ*RMST* expression patterns in these datasets. Several datasets including cancerous and non-cancerous tissues showed a highly variable range of *RMST* expression. Neuronal precursor cells had the highest mean *RMST* expression in the included datasets. Medulloblastomas samples show a *RMST* expression level lower than most glioblastomas samples but higher expression level than malignant melanoma samples. Notably, samples with the highest *RMST* expression levels were observed in the cancer cell line encyclopedia (CCLE) dataset. Grouped expression analysis according to cancer entities revealed that small cell lung carcinoma cell lines predominantly showed *RMST* overexpression in the CCLE dataset (data not shown).

4.5.2 Circ*ISPD* as SHH medulloblastoma biomarker

The *circs* pipeline revealed aberrant overexpression of circ*ISPD* in SHH MB for both datasets (Figure 19). Even though the overall expression of this candidate was lower than circ*RMST* expression, this biomarker candidate was still significantly overexpressed compared to the other groups in both datasets.

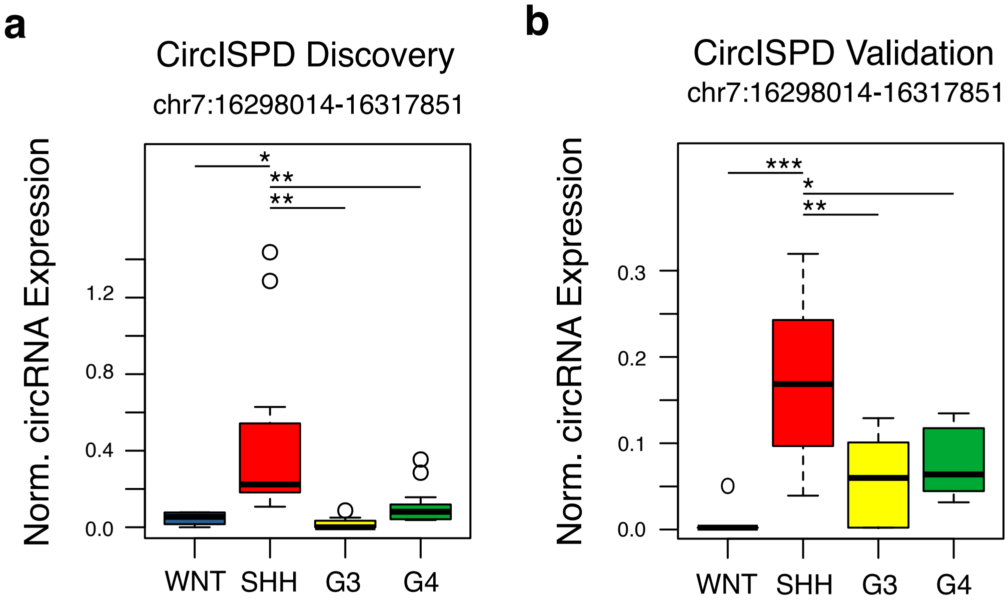


Figure 25: Circ*ISPD* normalized expression in circular RNA defined medulloblastoma groups in discovery (a) and validation (b) cohort with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001.

Notably, circ*ISPD* showed a significant variation of expression levels between the two cohorts. The mean circ*ISPD* level in SHH MB samples was 0.467 and 0.161 back-splice junction reads per million in the discovery and the validation cohorts, respectively. The discovery cohort data showed two outliers in the SHH group that expressed even higher levels of the candidate circRNA. Both datasets showed a partial overlap of the circ*ISPD* expression levels in Group 3 and Group 4 MBs, with SHH MB indicating worse biomarker potential for group determination compared to circ*RMST*. According to circbase, several circ*ISPD* isoforms have been detected in brain tissue, including the isoform identified in this study. The parental gene was investigated further to gain an insight into the functional importance of this locus. *Isoprenoid synthase domain containing (ISPD)* is a protein-coding gene on chromosome 7p21 (hg19 genomic coordinates: chr7:16127152-16460947) that has been associated with Walker-Warburg syndrome and muscular dystrophy by SNP-analysis[133,134]. This gene has been renamed to CDP-L-ribitol pyrophosphorylase A (*CRPPA*) and includes 10 exons. One non-coding antisense transcript is known.

4.5.3 Circ*EXOC6B* as Group 4 medulloblastoma biomarker

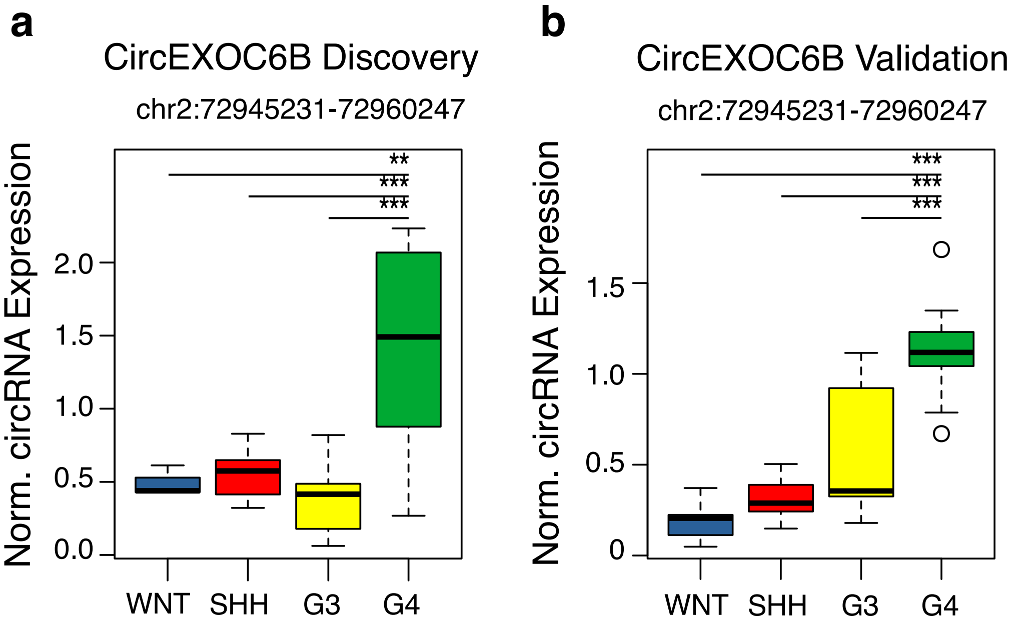


Figure 26: Circ*EXOC6B* normalized expression in circular RNA defined medulloblastoma groups in discovery (a) and validation (b) cohorts with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001.

In both datasets, circ*EXOC6B* was identified as the most promising biomarker of Group 4 MBs (Figure 19). However, it also showed a higher expression level in Group 3 MB samples in the validation cohort. Furthermore, increased circ*EXOC6B* expression levels in other non-Group 4 MB samples were comparable to those in Group 4 MB samples with lower circ*EXOC6B* expression. The linear *EXOC6B* transcript was expressed at relatively high levels (data not shown). Overall, the overlap in the circ*EXOC6B* expression of some non-Group 4 MB samples and the highly expressed linear isoform render this candidate less desirable for clinical application. Circ*EXOC6B* has been previously identified in brain samples according to circbase, and the exact isoform identified in our biomarker study has been found to regulate neural gene expression[135]. Furthermore, several genomic alterations of the parental gene have been implicated in intellectual disability[136(p6)]. The parental gene, Exocyst Complex Component 6B (*EXOC6B*) encodes a protein which is a part of the evolutionarily conserved exocyst, a multimeric protein complex necessary for exocytosis, which, in turn, is crucial for cell growth, polarity and migration[137(p6),138(p6)]. *EXOC6B* is located on the minus strand of chromosome band 2p13.2, precisely at hg19 genomic coordinates chr2:72403113-73053162. The gene consists of 28 exons, resulting in several alternatively spliced circRNAs.

4.5.4 Other medulloblastoma group-specific circular RNA biomarkers

Another set of biomarkers was identified to complement the candidates described above using an additional approach. Specifically, the two cohorts were combined into one cohort and fetal brain control samples were also included. As a result, more MB group biomarkers were detected (see Figure 27). Most of these additional candidates were expressed at a relatively low level, making them unsuitable for further investigation. Circ*PATJ* (Figure 27a) was a significant WNT-MB-specific circRNA, but showed highly variable expression in this group and a low overall expression. In addition, circ*PATJ* expression overlapped between

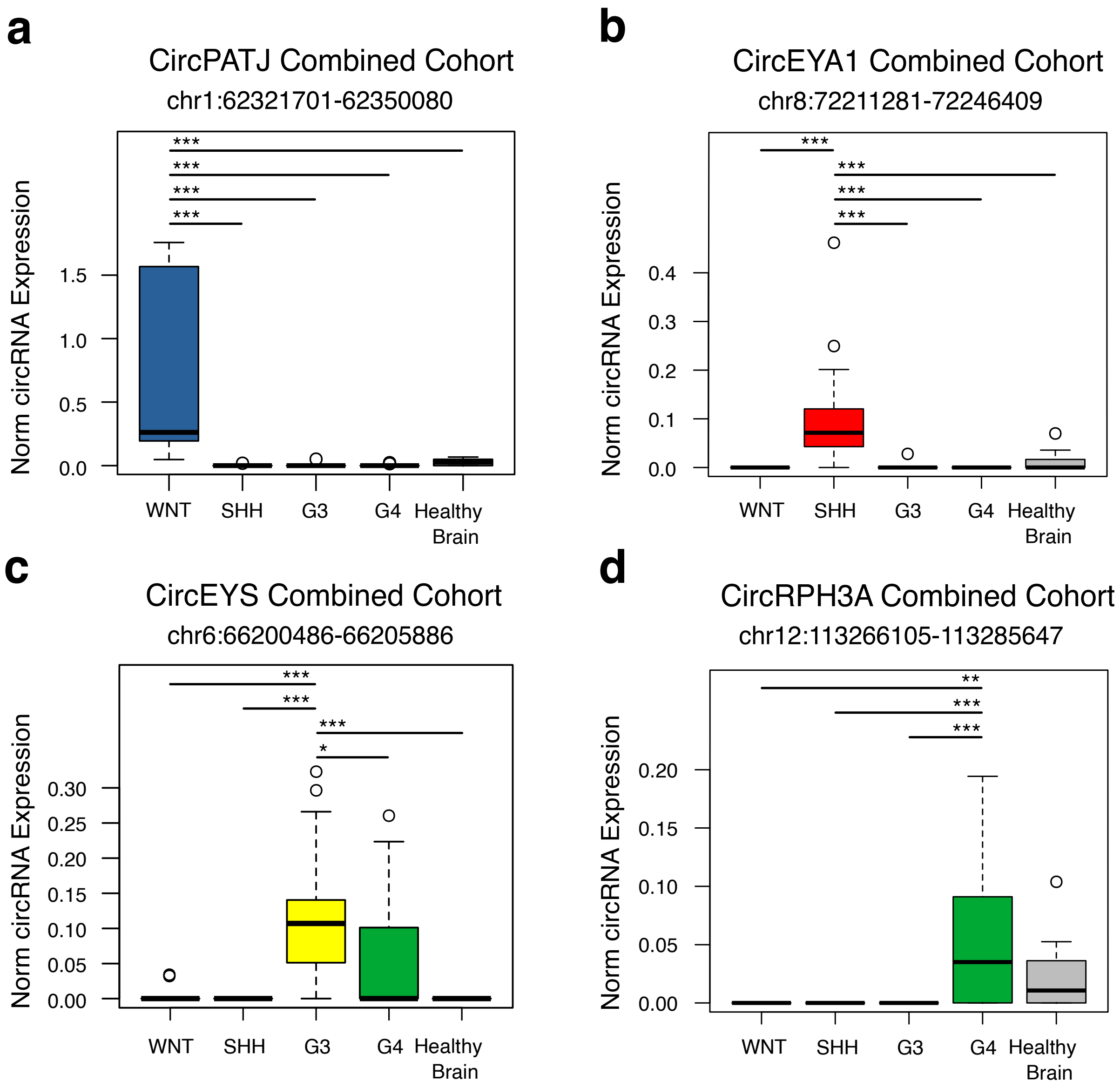


Figure 27: Circ*PATJ*, circ*EYA1*, circ*EYS* and circ*RPH3A* normalized expression in circular RNA defined MB groups in merged discovery and validation cohort combined with 12 fetal brain tissue controls with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001.

WNT MB and other samples from non-WNT MB groups, as shown in the plot. The SHH MB biomarker circ*EYA1* also had low overall but highly differential expression between the MB groups and healthy control samples. However, overlapping expression was seen in SHH MB and healthy control samples. The additional biomarker candidate circ*EYS* showed a different pattern: This circRNA was identified as a significant biomarker for Group 3 MB. However, based on the expression levels of this circRNA, no clear distinction could be made between Group 3 MB and Group 4 MB, due to the highly similar expression in these groups. One WNT MB outlier also had increased circ*EYS* expression comparable to Group 3 MB cases in the pooled cohort. Circ*RPH3A* showed an overlap between the median of Group 4 MB and the 3rd quartile of healthy control samples, making this biomarker unsuitable for clinical biomarker development. Overall, these additional biomarkers were mainly expressed at low levels and lacked a high specificity, as demonstrated by overlap of circRNA expression between groups including healthy controls samples. Therefore, these circRNAs should be deprioritized for future biomarker development.

4.6 Medulloblastoma-subtype-specific circular RNA biomarkers

To investigate the potential of circRNA as a biomarker for MB subtypes – not MB groups as previously discussed – the two cohorts were again visualized in a heatmap of the top 500 differentially expressed circRNAs. To discern subtypes however, the color annotation was shifted to show the DNA methylation-based subtypes of both cohorts.

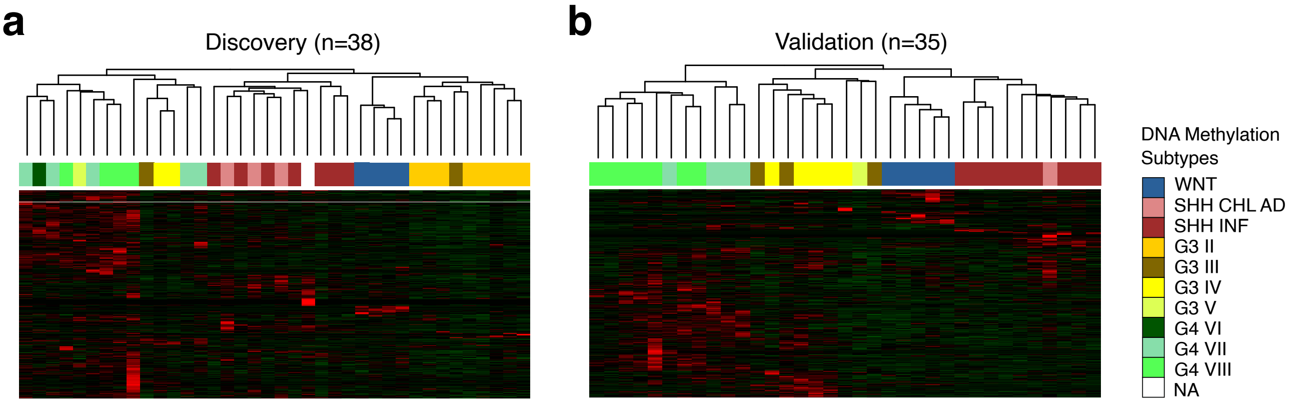


Figure 28: Heatmap of the top 500 differentially expressed circular RNAs in the discovery (a) and validation (b) cohorts. Clustering according to Pearson's dissimilarity, color according to DNA methylation subtype information (see legend on the right side).

The heatmaps shown here revealed a high degree of MB group clustering, but also a low degree of MB subtype separation (Figure 28a). The validation cohort, on the other hand, showed a clear separation between Group 4 subtype VIII and subtype VII. The clear WNT MB separation in both cohorts was the same, as expected, since all WNT MBs were also the same subtype, WNT. In both discovery and validation, the subtype separation of Group 3 MB subtypes was not exact according to the dendrograms. However, since the validation cohort included fewer subtypes than the discovery cohort – including samples with unknown subtype – the overall separation between MB subtypes was greater in the validation cohort. Since some MB subtypes were cohort-specific (only present in one of the two data sets, i.e. Group 4 VI only present in the discovery cohort), all further MB subtype investigation was made with the pooled cohorts, plus the healthy fetal brain controls, for a lack of better healthy brain tissue samples. Group 4 subtype VI is a good example of excluded subtypes; this thesis does not make claims based on a n = 1. On the heatmap annotation, the subtype allocation shows more details. SHH CHLD AD is present in only one sample in the validation cohort, but in three in the discovery cohort. Additionally, Group 3 type II is seen 8 times in the discovery cohort and is not seen in the validation cohort. Conversely, Group 3 IV is seen in the validation cohort seven times and three times in the discovery cohort. A more detailed subtype allocation overview is shown in Table 6.

Table 6: Comparison of DNA methylation subtypes in the discovery and validation cohorts.

subtype	discovery	validation	sum	subtype discovery
NA	1	0	1	no
WNT	4	5	9	Not useful
SHH CHL AD	3	1	4	yes
SHH INF	7	9	16	yes
Group 3 II	8	0	8	no
Group 3 III	2	3	4	yes
Group 3 IV	2	5	7	yes
Group 3 V	1	1	2	no
Group 4 VI	1	0	1	no
Group 4 VII	5	4	8	yes
Group 4 VIII	4	7	11	yes
Sum	38	35	73	

To further investigate the potential for DNA methylation subtype biomarkers in the circRNA data, only subtypes that are spread throughout both data sets and provide a total number of samples greater than 4 were investigated. For this analysis, an ANOVA was carried out, followed by a TukeyHSD as the post-hoc test. All mentioned p-values are adjusted p-

S

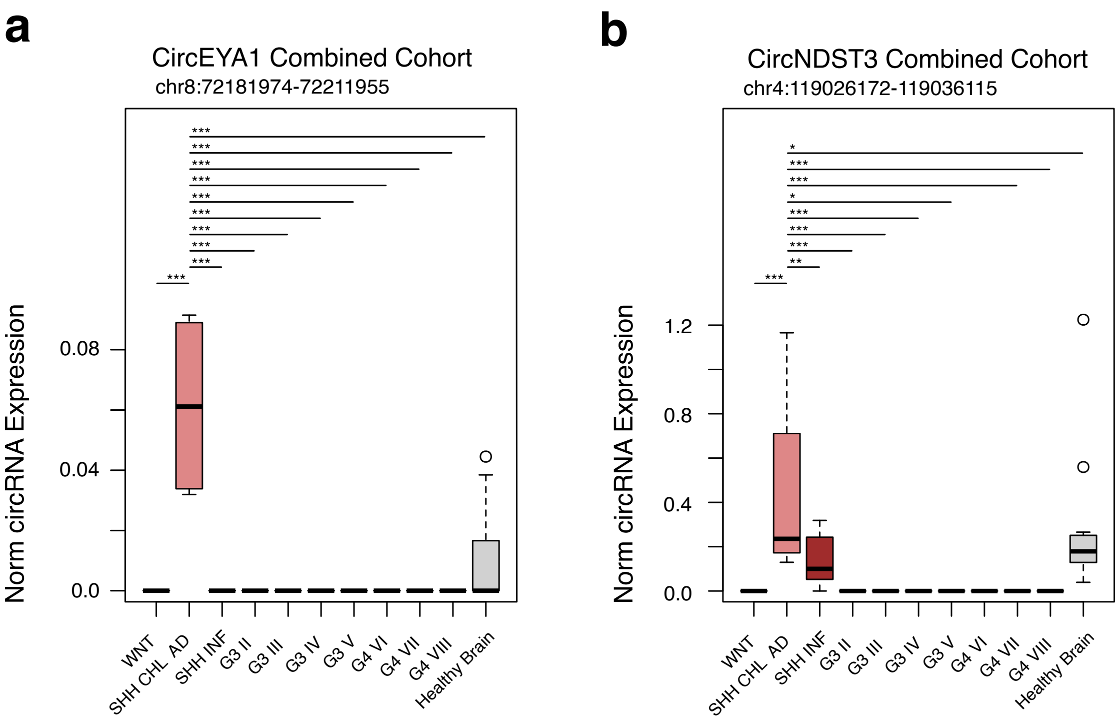


Figure 29: SHH CHLD AD subtype circular RNA biomarkers. Based on DNA methylation subtypes and combined cohort with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001.

The circRNA biomarker candidates for SHH CHLD AD shown in Figure 29 displayed significant differences in circRNA expression between DNA methylation-based subtypes. Both candidates showed expression levels overlapping with the healthy brain samples included in this plot. CIrc*NDST3* also showed an overlap of expression with the other SHH subtype, SHH INF. Due to this amount of overlap between different subtypes, these candidates were not followed up in further studies.

4.6 Validation of *in silico* prediction based on circs workflow

To evaluate the identified circRNA biomarker candidates and all circRNA species detected by *circs*, a number of verification methods were used. First, we used *in silico* methods to overlap the detected circRNA coordinates with circRNAs from other databases, then the circleseq protocol was used to evaluate the number of RNAseR-resistant circRNA species, indicating the true detection rate in the datasets shown here.

4.6.1 Database-aided validation of circRNAs detected with *circs*

The first and foremost *in silico* validation method is the built-in overlap with circbase, the first circRNA database including human data from several experiments. To assess the overlap with circbase, the total amount of unique circRNA species was divided into two parts: the circRNAs with a circbase id attached and the circRNAs without one.

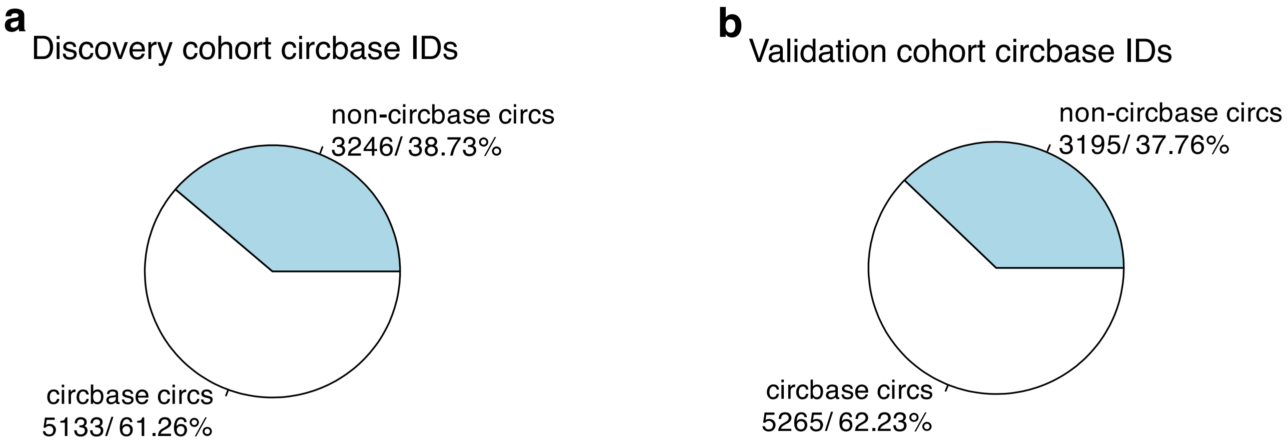


Figure 30: Circbase-based circular RNA validation of discovery cohort (a) and validation cohort (b).

The overlap with circbase showed a similar percentage of circRNAs confirmed in both datasets. However, this database is neither exhaustive nor focused on samples of malignant brain tumors. To correct this missing overlap in sample number and tissue of origin, this overlap was repeated with a larger, more exhaustive and cancer-focused circRNA database, MiOncoCircDB.

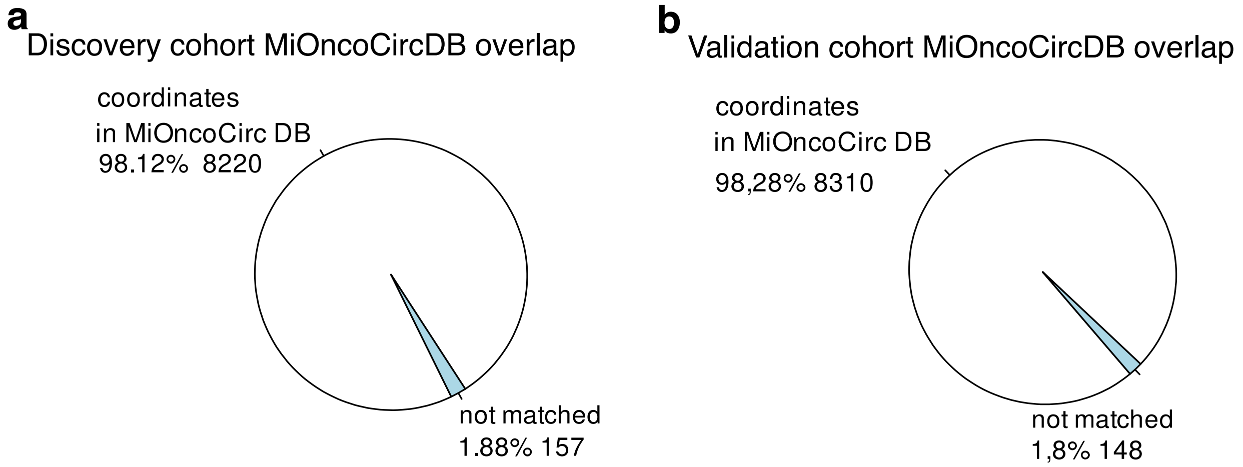


Figure 31: MiOncoCircDB coordinates overlaps with discovery (a) and validation dataset (b). LiftOver was used to translate the hg19 coordinates of both datasets into hg38 coordinates, failed liftover coordinates were excluded from this analysis.

Since MiOncoCircDB is not a manually curated circRNA database, some of its data could have quality control issues, so circRNAs of both datasets were again overlapped with all curated circRNA databases available at date of analysis. The striking similarity between the discovery and validation cohorts again resulted in similar overlap results.

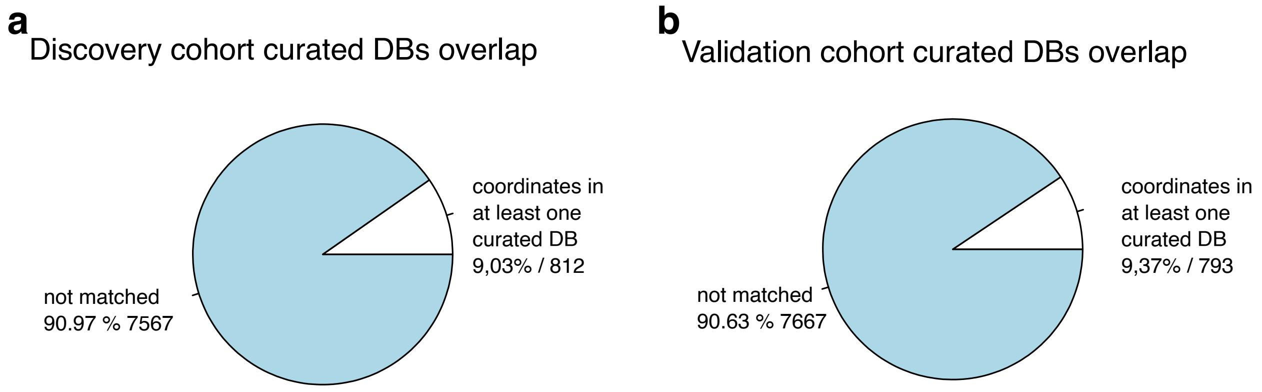


Figure 32: Overlap of discovery (a) and validation (b) cohort with all curated data bases.

The comparatively small overlap of both cohorts with all curated databases, seen in Figure 32, could be explained by the nature of these databases themselves: the content was manually curated, which leads to a smaller total number of circRNAs included in the databases compared to their non-manually curated counterparts.

4.6.2 Quantitative circular RNA *in silico* specificity determination

To investigate the specificity of the top 3 chosen biomarker candidates, circ*RMST*, circ*ISPD* and circ*EXOC6B*, data from the largest cancer-focused circRNA database was used and overlapped with the data of the validation cohort for reference.

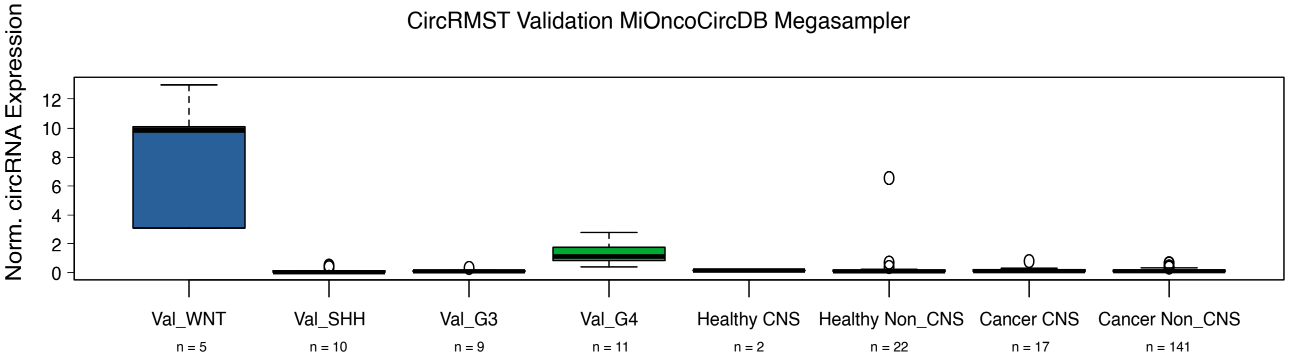


Figure 33: Circ*RMST* in validation cohort and MiOnciCircDB data divided into four categories: Healthy_CNS, Healthy_Non_CNS, Cancer_CNS and Cancer_NON_CNS. All MiOncoCircDB samples with detected circ*RMST* were included in this plot.

The staggering amount of circ*RMST* specificity in this dataset was disturbed by one sample in the Healthy_Non_CNS group of the MiOncoCircDB data, a healthy control prostate sample from non-neurologic tissue of origin. This assessment was continued for the remaining two circRNA biomarker candidates, circ*ISPD* and circ*EXOC6B*.

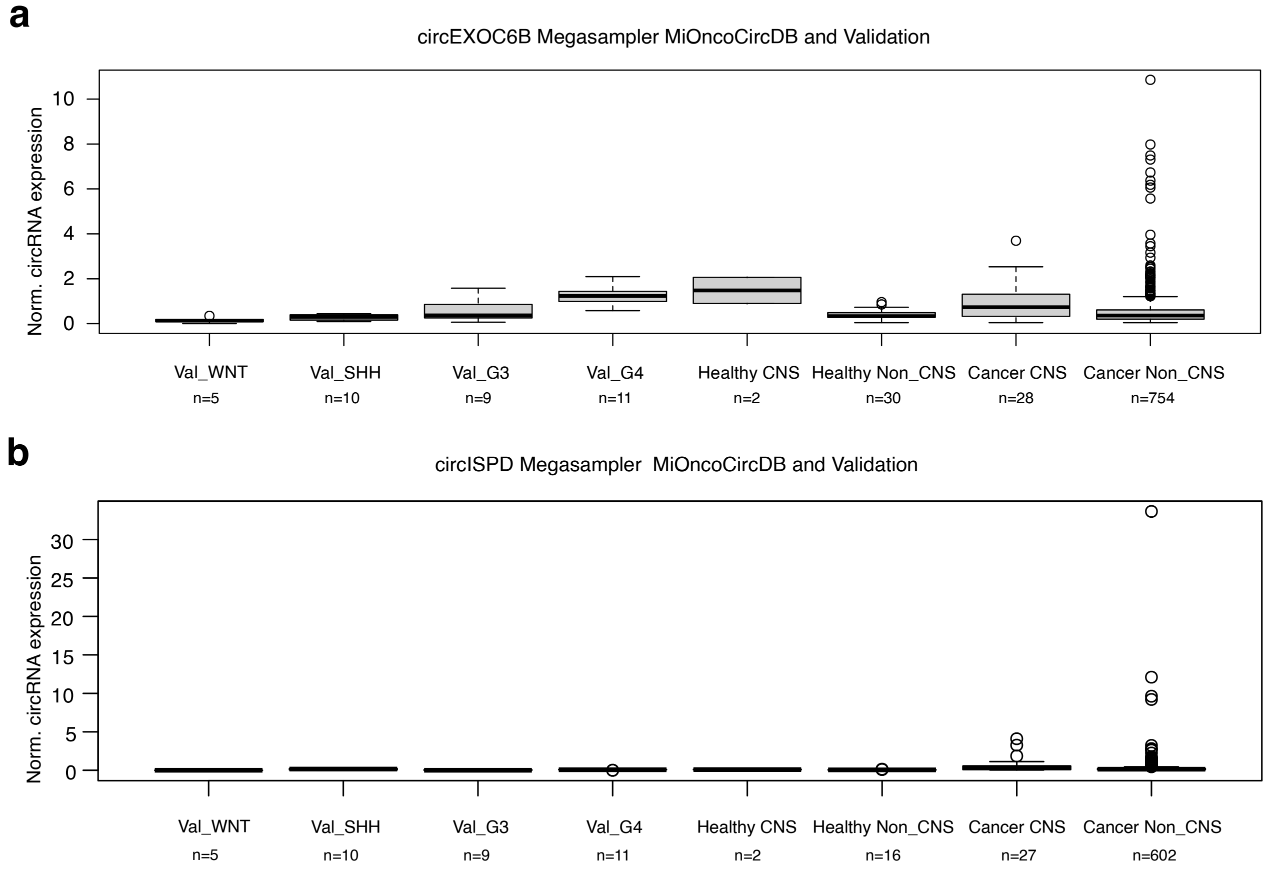


Figure 34: MiOncoCircDB megasamplers for circ*EXOC6B* (a) and circ*ISPD* (b) in validation and MioOncoCircDB. N numbers refer to total samples in the validation cohort and circular RNA positive samples for the respective circular RNA in each plot.

These plots show a wide variety of cancerous and non-cancerous tissues in MiOncoCircDB, expressing circ*EXOC6B* and circ*ISPD* in much higher levels than in their respective biomarker groups. The plot a depicts a high expression of circ*EXOC6B* in validation Group 4 MB (as previously eluded to), but also a comparably high mean expression in both healthy central nervous system samples and even higher outliers in both tumor categories, inside and outside the central nervous system. Next, an orthogonal method was used to further assess the biomarkers and all detected circRNAs in both datasets.

4.6.3 Circleseq based circular RNA validation in an *in vitro* model

To demonstrate *circs* as a valid circRNA detection pipeline delivering reliable and precise results, the Circleseq protocol was used. The total circRNA signal detected for each condition was measured to confirm a successful RNAseR treatment of the three cell lines. The results can be seen in Figure 35.

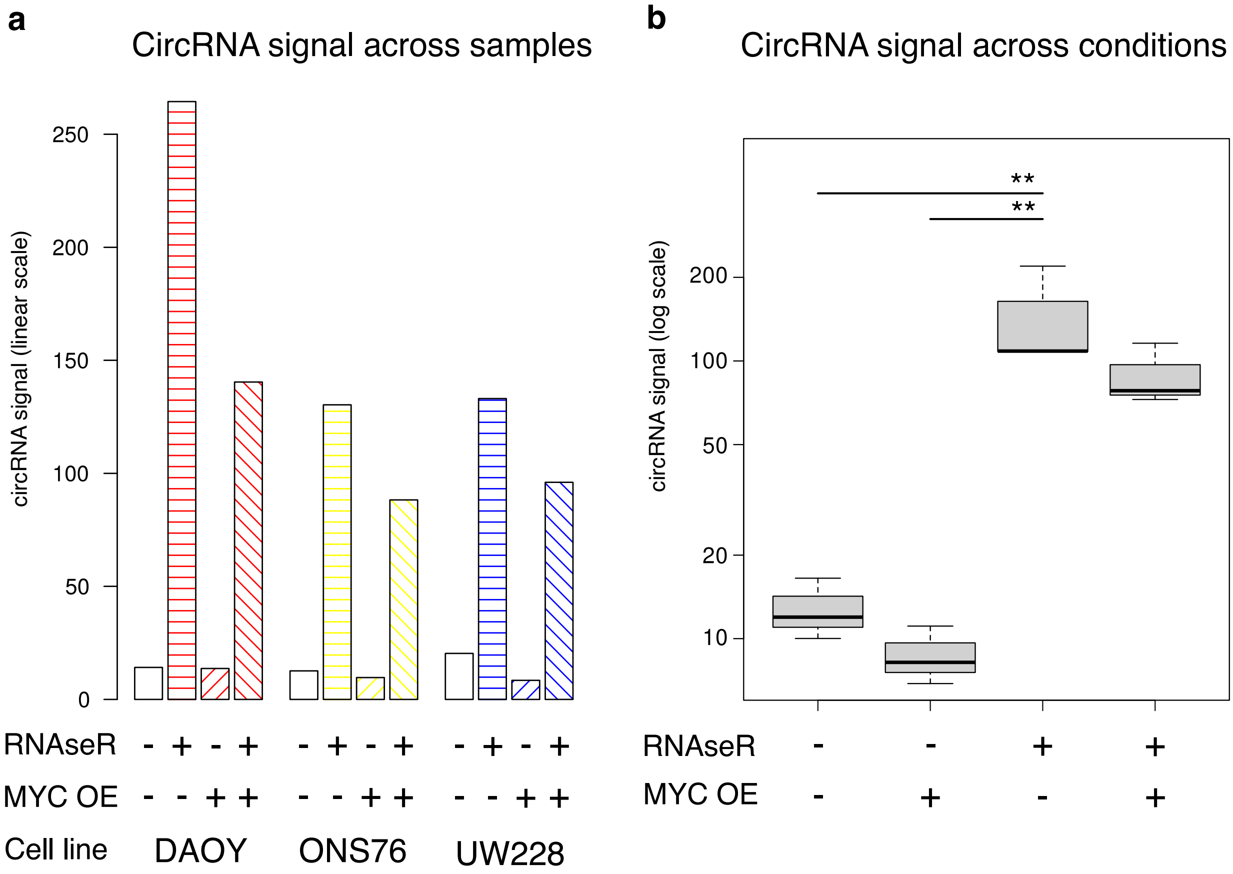


Figure 35: Normalized circular RNA signal. a) for each sample, ordered by cell line and condition. b) across 4 conditions in 3 **medulloblastoma** cell lines combined**.** RNAseR= sample treated with RNAseR. ** = Tukey HSD p-value adjusted <0.01. MYC OE= MYC Overexpression.

Figure 35 shows a substantial circRNA enrichment across all three cell lines upon RNAseR treatment, as expected. Additionally one can observe that the *MYC* samples (samples overexpressing *MYC*) show a decreased circRNA signal compared to the control (ctrl) samples. Since this is an isogenic model, outside of analysis or experimental errors, this can be traced back to the *MYC* overexpression and its effects. Another observation from this plot is that each cell line had a different level of baseline circRNAs present, which was increased with a sample-specific factor for each of these cell lines (see Table 7 for circRNA signal details).

Table 7: Circular RNA signal across all *MYC*/RNAseR samples. Circ signal decrease calculated in percent of baseline, *MYC* OE effect calculated based on x times baseline.

Sample	**circ signal ctrl untreated**	**circ signal MYC untreated**	**circ signal ctrl RNAseR**	**circ signal MYC RNAseR**	**circ signal decrease unpon MYC OE - no RnaseR**	**circ signal decrease unpon MYC OE - RNAseR treatment**	**RnaseR circ signal amplification: ctrl**	**RnaseR circ signal amplification: MYC OE**
DAOY	11.95	11.1	219.55	115.81	7.11	47.25	17.37	9.43
UW229	16.5	6.89	107.91	78.12	58.25	27.6	5.54	10.34
ONS76	10.02	8.22	108.53	72.62	18	33.09	9.83	7.84

In addition, Figure 35b shows the same data sorted into four conditions. Here the previously described trend of lower circRNA signal as a result of *MYC* overexpression was visible. Table 7 shows a global circRNA signal decrease upon *MYC* overexpression, in untreated and RNAseR treated samples, with RNAseR-based numbers being higher except for the UW228 cells. Additionally, the RNAseR-treated samples showed a higher level of circRNA signal. The only significant differences between these treatment groups were between the RNAseR samples without *MYC* overexpression and both non-RNAseR-treated groups. The cell line specificity of the resulting circRNA signal observable in Figure 35a can also be observed in the heatmap of the top 10% differentially expressed circRNAs across this dataset.

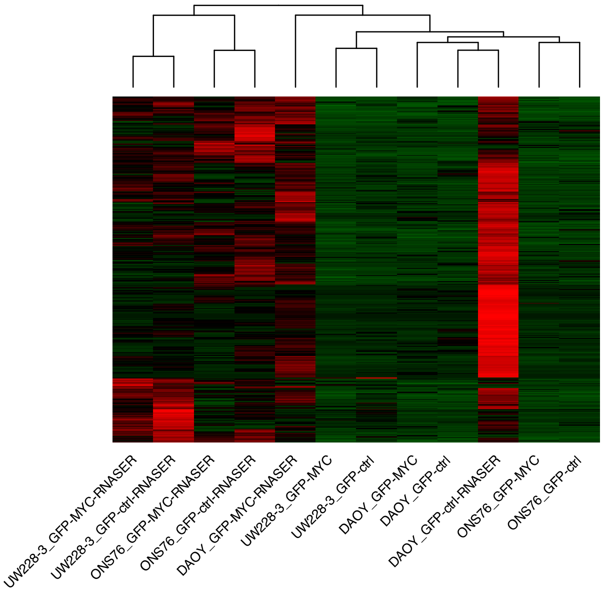


Figure 36: Heatmap of top 500 differentially expressed circular RNAs across the *MYC*/RNAseR dataset. Normalized circRNA data. Clustering based on Pearson's dissimilarity, high circRNA expression = red, low circRNA expression = green.

In the heatmap depicted in Figure 36, the strong circRNA signal increase is observable throughout all RNAseR-treated samples (red = high signal). Generally, the RNAseR-treated sample of one cell line clustered closely to the RNAseR-treated *MYC*-overexpressing sample, pointing further to strong circRNA signal increase upon RNAseR treatment. Since the non-treated samples had a much lower overall signal (Figure 35), these clustered together, with the *MYC*-overexpressing sample being next to its relative control. The DAOY control cells that were RNAseR treated showed a clear difference: this sample clustered more closely to the same cell line in other treatment conditions rather than other cell lines in the same treatment conditions.

4.6.4 *In silico* validation of experimental data

As before, the first step in evaluating the *MYC*/RNAseR dataset was the *in silico* validation. The included circbase overlap was performed as a first validation effort (Fig. 37).

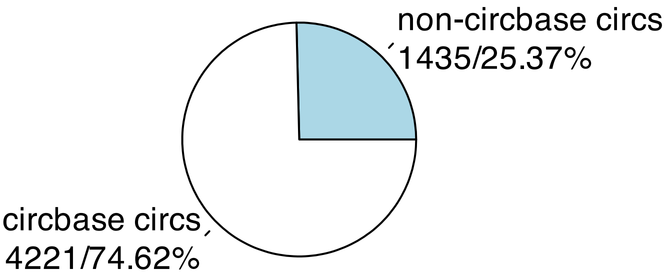


Figure 37: Circbase - based overlap of circular RNA candidates included in the whole *MYC*/RNAseR dataset.

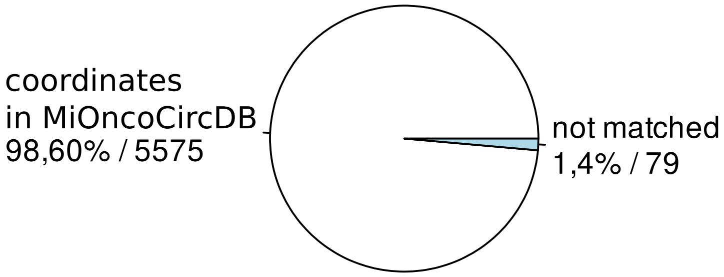


Figure 38: *MYC*/RNAseR samples and MiOncoCIrcDB overlap. All in the dataset included circular RNAs were used for this plot.

This pie chart shows an elevated (in comparison to the discovery and validation cohorts) circbase validation ratio of 74.62%. This included the circRNA species from the non-treated and RNAseR-treated samples. For the next validation step, the circRNAs were overlapped with MiOncoCircDB data, leveraging the more cancer-specific circRNAs. This plot, like the previous one, shows a relative increase of circRNAs identified in the database (Fig. 38). This could be a sign of the dataset including a smaller number of circRNAs, or a generally higher number of true positive circRNAs. The next *in-silico* validation step of the *in-vitro* validation data was the overlap of circRNA detected by circs with the curated databases.

The next step in this analysis was the validation of the detected circRNAs. All circRNAs also detected in the RNAseR treated samples were considered true positives, all circRNAs not detected in the with RNAseR treated one false positive. All *in-silico* validation rates calculated here are summarized in Table 8.

Table 8: *In silico* validation rate of the three MB circRNA datasets included in this thesis.

Dataset	circbase overlap	MiOncoCircDB overlap	Curated DBs overlap
Discovery	61.26%	98.03%	9.03%
Validation	62.23%	98.28%	9.37%
MYC/RNAseR	74.62%	98.60%	12.90%

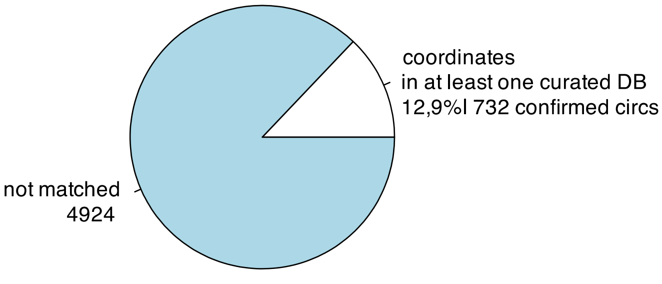


Figure 39: *MYC*/RNAseR samples and curated databases overlap.

The overall trend of an increased validation rate in the *MYC*/RNAseR data set compared to the discovery and validation cohorts across all databases used continued and can be observed in Table 8.

The *MYC*/RNAseR dataset was then used to calculate the rate of RNAseR–stable circRNAs across all cell lines used.

4.6.5 RNAseR validation rate assessment

To calculate the overall RNAseR validation rate, circRNAs detected in samples that were not treated with RNAseR were compared to those that were. In this approach, one cell line can confirm previously unconfirmed circRNAs from another cell line.

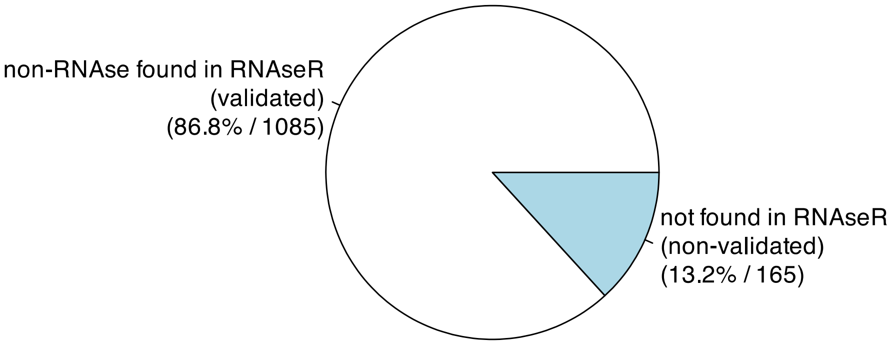


Figure 40: RNAseR stable circRNAs found in the dataset relative to all non-RNAseR circRNA species across all three used cell lines.

The pie chart in Figure 40 shows a relative validation rate of 86.8% in the combined *MYC*/RNAseR samples. Overall, 1085 circRNAs could be validated. Another form of visualizing the same result can be seen below as a Venn diagram, additionally visualizing the circRNAs that were exclusively detected in the not RNAseR treated samples (Fig. 41).

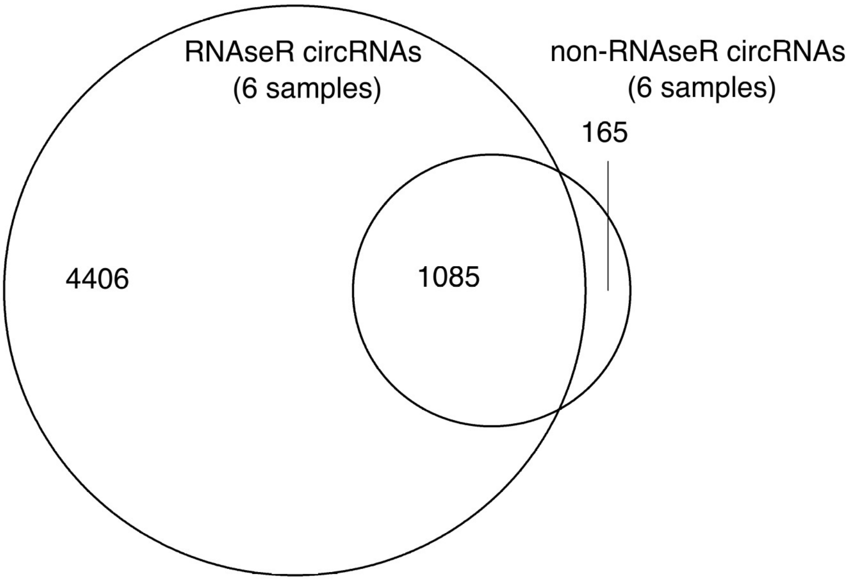


Figure 41: RNAseR and non-RNAseR circular RNAs overlap in the *MYC*/RNAseR dataset.

Clearly, the larger number of individual circRNAs can be found in the RNAseR-treated samples. However, 165 non-RNAseR circRNAs (13.2%) could not be validated using this approach. Figure 41 also shows one key shortcoming of this data: there are many RNAseR circRNAs that were not detected in the non-RNAseR samples. There are multiple possible reasons for this that will be evaluated in the Discussion. Furthermore, circ*EXOC6B* and circ*ISPD* were detected in RNAseR and non-RNAseR treated samples in the *MYC*/RNAseR dataset, but circ*RMST* was not detected. To maximize the insight gained from the *MYC*/RNAseR dataset, the previously shown discovery and validation cohorts were overlapped with the complete *MYC*/RNAseR cohort in Figure 42a, and with only the RNAseR-stable circRNAs included in this dataset in Figure 42b and Figure 42c for discovery and validation, respectively.

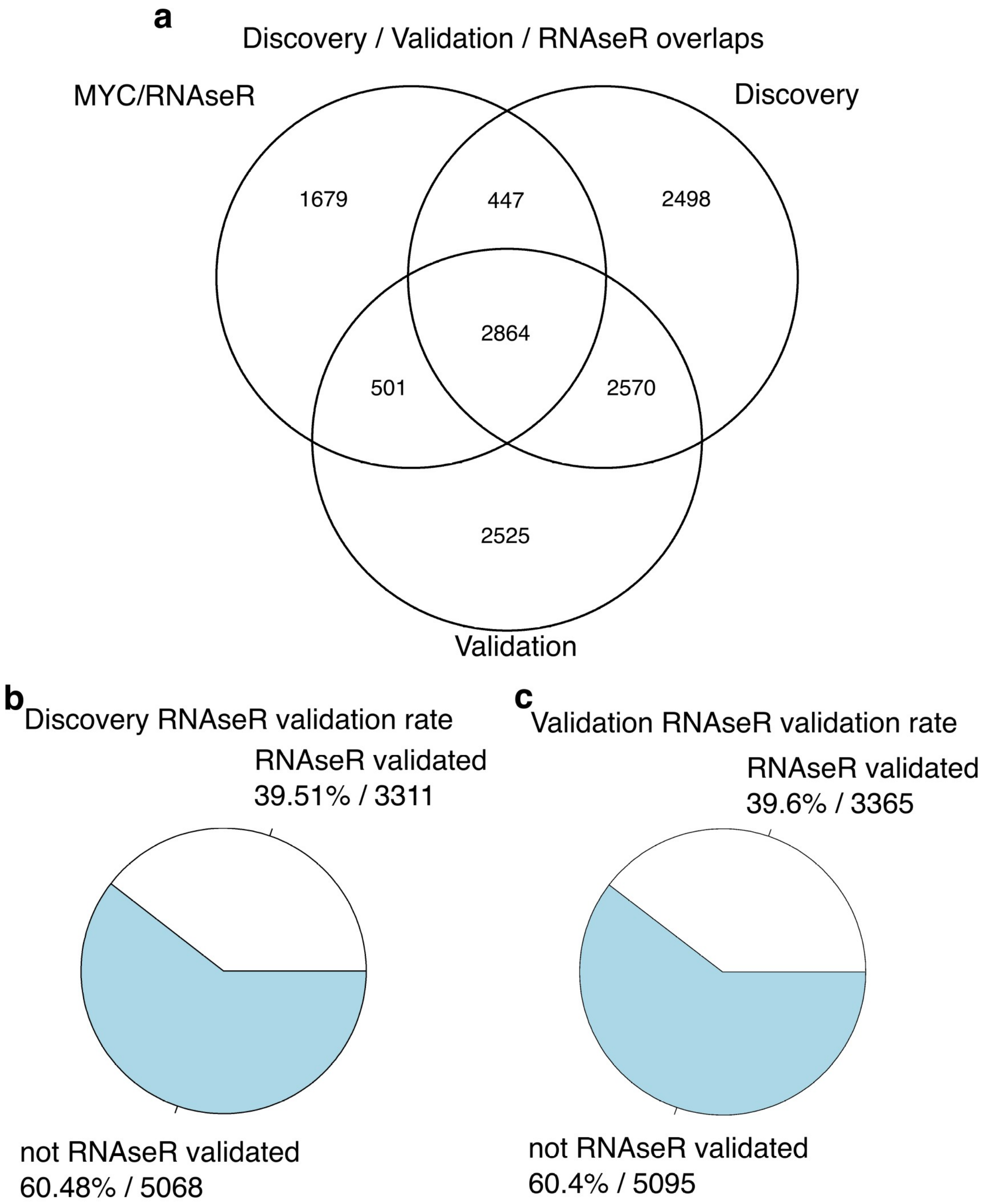


Figure 42: *MYC*/RNAseR circRNA overlaps with discovery and validation cohort. Overlaps of complete data (a) and Overlap of discovery (b) and validation (c) cohort with RNAseR-treated samples respectively.

The validation rates of the discovery and validation cohorts depicted in this figure must be taken as estimates; the RNAseR-treated samples are not able to accurately describe the whole intertumoral heterogeneity of any typical MB. Additionally, the cell lines were not representative of all four MB groups mentioned in this thesis. Further, the full dataset was not as comprehensive as the discovery or validation cohorts, based on the much smaller number of included samples (three cell lines grown *in vitro* versus 35 or 38 patient-derived samples).

4.6.6 Cell-line-specific RNAseR validation rate

To further elucidate the circRNA detections made in the paired treated and untreated samples and to define cell-line-specific RNAseR validation rates, the data set was split for each cell line. Here, two samples are compared to each other: RNAseR treated versus the same condition (*MYC* or ctrl) untreated. The results are shown in Figure 43.

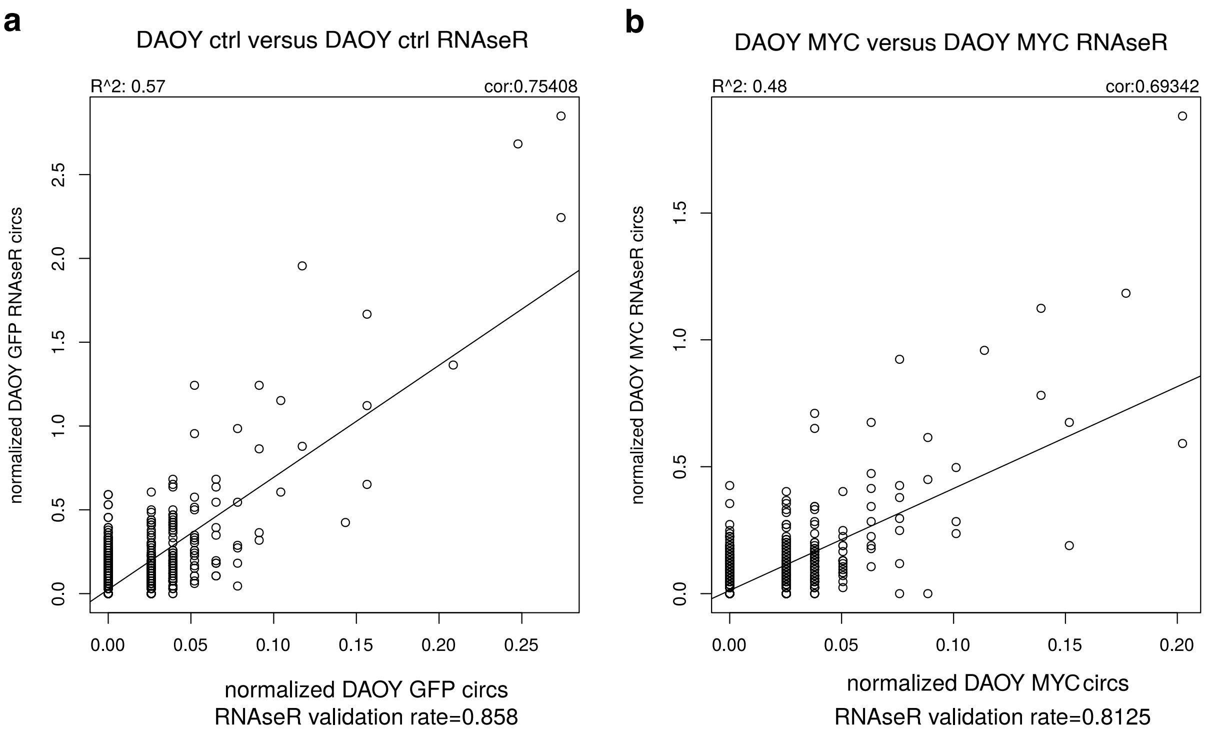


Figure 43: DAOY-specific RNAseR validation rate. R^2 based on a linear model through all data points.

Figure 43 shows a relative validation rate loss between ctrl and *MYC* samples in DAOY MB cells of 4.6%. The general correlation of circRNA quantification between the treated and untreated sample is higher in the ctrl sample.The axis on Figure 43a shows a 10-fold increase upon RNAseR treatment, hinting at the signal amplification rate of this RNAseR treatment. The untreated DAOY sample has the highest single-sample validation rate, 1% less than the overall validation rate considering all cell lines where each RNAseR sample

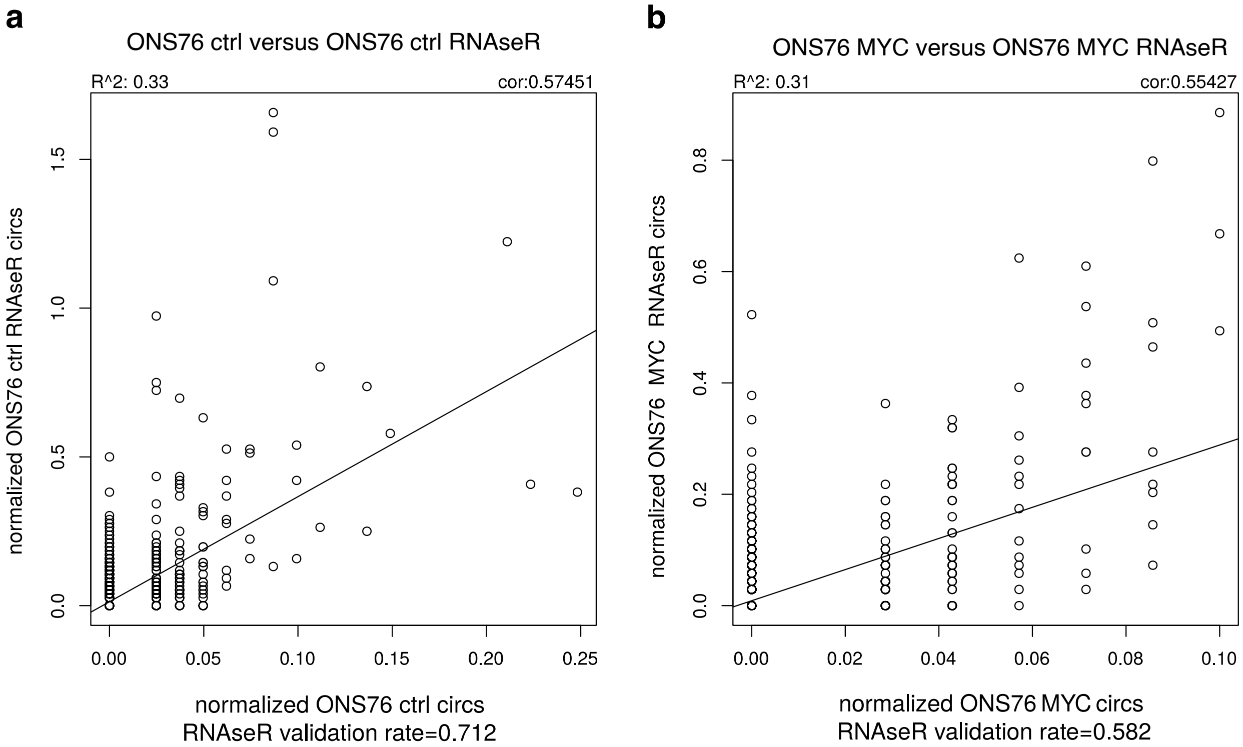


Figure 44: ONS76-specific RNAseR validation rate. R² based on a linear model through all data points.

can confirm non-treated circRNA detection in another cell line..

Figure 44 shows a comparatively lower validation rate, also decreasing for the *MYC* samples by 13.9%. The correlation between RNAseR-treated and -untreated quantifications is much lower compared to the DAOY samples shown above.

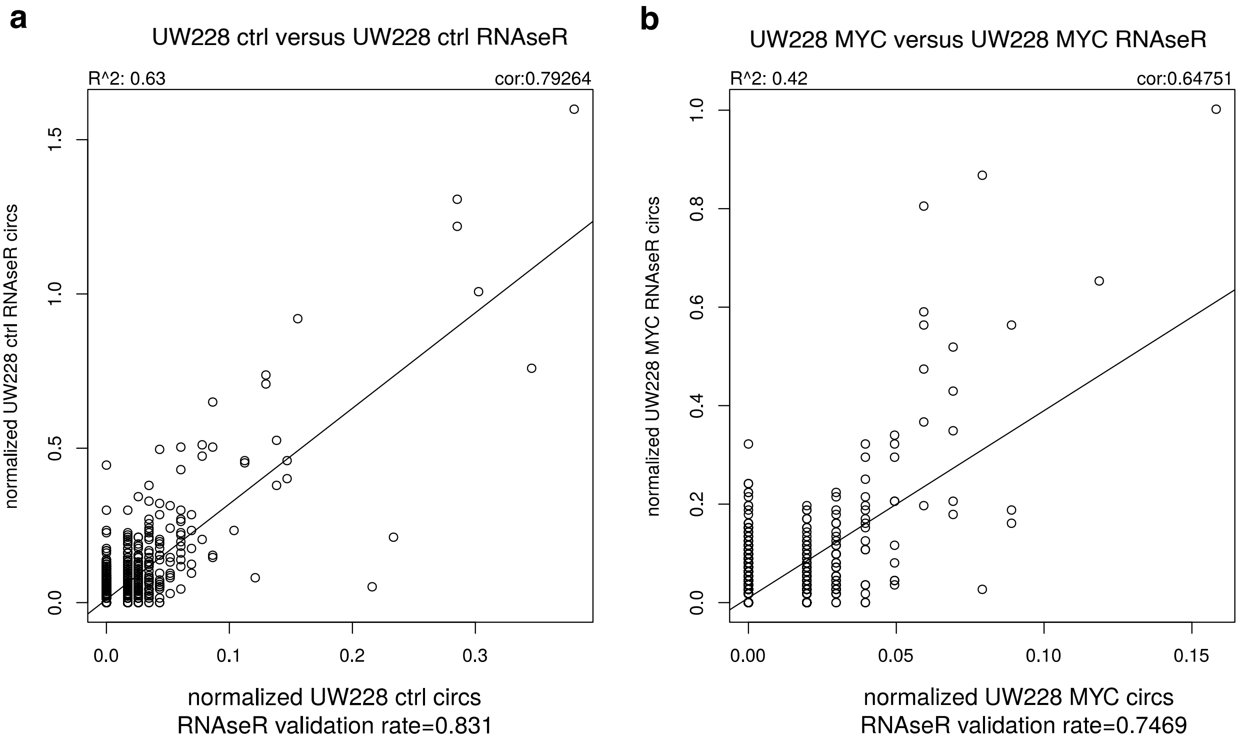


Figure 45: UW228-specific RNAseR validation rate. R^2 based on a linear model through all data points.

Figure 45 shows an intermediate result for the UW228 MB cell line: the validation rates in both conditions are lower compared to DAOY cells, yet higher than the ONS76 cells. However, for the non-*MYC* sample, the correlation between the RNAseR-treated and RNAseR-untreated samples was the highest in all comparisons shown here. Arguably, the cell-line-specific RNAseR validation rates were more precise compared to the samples including a pooled overlap of circRNAs.

4.6.7 Circular RNA pertubations in the *MYC*/RNAseR dataset

To evaluate the direct effect of *MYC* overexpression in MB cell lines on circRNA detection, a volcano plot was created showing two dimensions of the *MYC*-specific circRNA pertubations: fold change and p-value. For this, all cell lines were pooled again, this time based on their *MYC* status, yielding two groups with 6 samples in each.

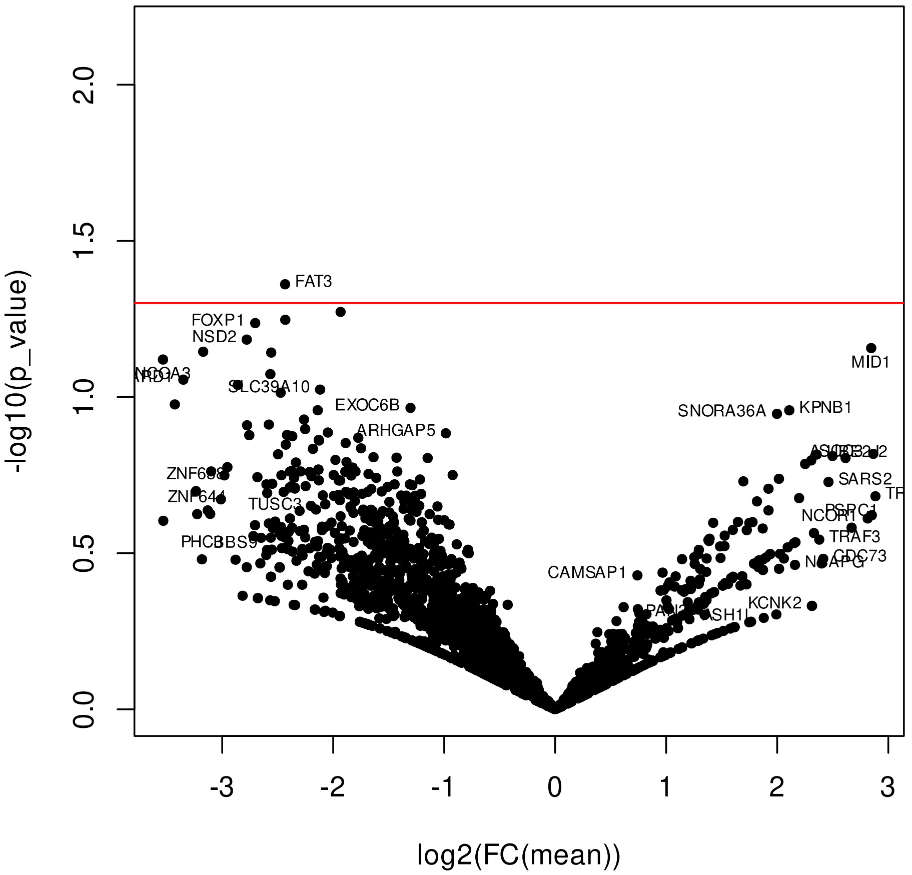


Figure 46: Volcano plot of *MYC* overexpression specific circular RNA changes. red line signals statistical significance. the parental gene names of some circular RNAs are next to its datapoint.

The differentially expressed circRNAs shown here spread in both directions of the "volcano". All data points to the right side are increased upon *MYC* overexpression, and the opposite is true for the left side. Only one circRNA species (circ*FAT3*) was significantly perturbed in the ctrl samples compared to the *MYC* overexpressing cells. Generally, more circRNAs were downregulated upon *MYC* overexpression, indicated by the higher number of circRNAs on the left side of the volcano plot.

5 Discussion

In the last 10 years, considerable progress has been achieved in the still relatively young circRNA research field[49-63]. However, there are still no circRNA-based biomarkers in daily use today. Given the emerging field of circRNAs as biomarkers for disease[104,91,86-88], this thesis aimed to identify and validate the first circRNA-based biomarkers in MB, a highly malignant and biologically heterogeneous pediatric cancer[14-18].

To reliably detect and quantify circRNAs from RNA-Seq data, *circs* was established, a pipeline enabling researchers to use the output of three circRNA detection algorithms. Using this pipeline in two non-overlapping MB cohorts, circRNAs were detected and quantified, revealing MB-group-specific circRNA signals in both cohorts. In the analysis of the top 500 differentially expressed circRNAs in both cohorts, it was revealed that clustering the circRNA data can classify MB samples into their respective MB groups almost as precisely as SNF, a much more costly and data-demanding method. Additionally, Group 3 MB expressed the lowest number of circRNAs in both cohorts, suggesting that the total circRNA signal observed in each sample could potentially be relevant regarding the underlying tumor biology or the clinical course of the patient.

The enrichment of circRNAs in brain tissues described in the literature and its relative depletion in MB was confirmed[57,61]. A stark contrast was observed in the circRNA signal between both MB cohorts and healthy fetal brain tissue. This needs to be taken with a grain of salt, because the healthy fetal brain tissue samples were sequenced as part of the ENCODE project much more deeply than both MB cohorts and on different platforms. The strong comparability between the two MB cohorts remained and illustrates the stability of circRNA-based biological classification across MB data sets.

Next, reproducible MB group-specific circRNA biomarker were revealed using both cohorts ('pan-cohort'). Circ*RMST* was the most highly expressed circRNA in WNT MB, circ*ISPD* in SHH MB and circ*EXOC6B* in Group 4 MB. Group 3 MB did not reveal any group-specific significantly upregulated circRNA. This might be due to the overall lower number of circRNAs in Group 3 MB observed across both cohorts. From all pan-cohort circRNA biomarkers found, the most highly expressed circRNA was identified as the “candidate” for its group, due to practical concerns to be outlined later. The circRNA biomarker candidates defined in this thesis, circ*RMST*, circ*EXOC6B* and circ*ISPD*, have several interesting shared properties. All of these candidates have been found in other datasets of circRNA expression, such as circbase. The parental genes of these candidates also show some common properties. They have all been investigated in relation to abnormal neuronal development and/or cancer[132,133]: *RMST* (rhabdomysarcoma-associated transcript) has this property already included in its name; *ISPD* is associated with Walker-Warburg syndrome (one key symptom being mental retardation and seizures); and *EXOC6B* has been found to play a role in intellectual disability[136]. However, none of these have been previously linked to medulloblastoma, neither in circular nor in linear forms.

The defined circRNA biomarker candidates were then validated in a cancer-focused circRNA database, MiOncoCircDB. Here, each of the three candidates was first confirmed to be present in the dataset. Secondly, the expression of the specific candidate was compared to all samples in this database, sorted into four distinct sample groups based on their tissue of origin: healthy_CNS, Healthy_non_CNS, cancer_CNS and cancer_non_CNS. This was used as an additional indicator of the candidates’ specificity throughout the whole body in states of health and disease. Here, circ*RMST* showed a high degree of specificity, with only one healthy prostate sample overlapping with the WNT MB circ*RMST* levels in the validation dataset. However, the other biomarker candidates were shown to be less specific this way; for circ*ISPD* and circ*EXOC6B*, higher expression values were found in CNS and non-CNS tissues, making neither sufficiently MB-specific, and thus both are not promising candidates for further circRNA biomarker investigations.

In a next approach, both cohorts were merged to identify DNA methylation-based subtype biomarkers. Here, the small number of samples in some subtypes made the search for biomarkers challenging, but some subtypes were identified as having specific circRNA biomarkers. However, these findings are based on small sample sizes, arguably lowering their value. Additionally, this approach did not yield highly specific biomarkers; the best subtype circRNA biomarkers presented overlapped with other subtypes of the same MB group.

To validate all circRNA species detected in the discovery and validation cohorts, first the circbase data included in *circs* was used to confirm circRNAs already published in other studies. For the same target, a number of manually curated databases were also used, and finally, MiOncoCircDB. All three *in silico* validation methods showed a slightly increased number of validated circRNAs in the validation cohort compared to the discovery cohort.

To further validate the identified circRNA biomarkers and the established pipeline, the Circleseq protocol was used. Here, three well-established MB cell lines (ONS67, UW228 and DAOY) were treated to overexpress *MYC*[42], a risk-indication for MB patients known from clinical data. The three cell lines, with and without *MYC* overexpression, were treated with RNAseR, an enzyme that digests linear RNAs but not most circRNAs in a given sample. This dataset, called *MYC*/RNAseR, was used to a) validate ~40% of all circRNAs detected in the validation and discovery cohorts and b) show the circRNA true detection rate of circs to be 86.8% in this data set, when all samples are used at once and the maximum cell-line-specific validation rate is 85.8%. Even though the three MB cell lines lack heterogeneity found in the discovery and validation cohort (not all four MB groups were represented), the achieved circRNA validation rate is surprisingly high and consistent. This relative validation rate can be seen as an indicator of missing sequencing depth in the cell lines, indicated by the high number of circRNAs from the RNAseR samples not found in the corresponding untreated samples. These circRNAs did not emerge due to the RNAseR treatment. They were supposedly already in the untreated samples, but only the RNAseR-based enrichment made these “visible”, essentially identifying one key argument for 86.8% not being the exact number of true positives in all three datasets. To evaluate this hypothesis, these samples would need to be sequenced again, but with a much higher amount of input material and a higher sequencing depth. However, the own findings are a good indication and in line with other publications that utilized the Circleseq protocol to identify the true-positive rate of their respective circRNA detection tools[49,100].

During this project, a pair of matched FFPE (formalin fixed paraffin embedded) samples – commonly used for routine histological studies and long-term sample storage and snap frozen samples of glioblastoma were analyzed as well (Supplementary Figure 47). The key finding here was that highly expressed circRNAs detected in the snap frozen tissue could also be found in the corresponding FFPE sample, albeit with a relative signal loss of ~ 75%. This preliminary data suggests that elevated circ*RMST* expression could be determined in RNA extracted from FFPE samples to identify WNT Mbs(similar to 107).

Additional data not shown in this thesis include other exclusion reasons for circ*EXOC6B*. The lin-to-circ ratio (the fraction of linear transcript versus the number of backsplice junctions found in the same data) was ~1 for circ*RMST*, indicating complete or almost complete circular isoform dominance, while circ*EXOC6B*'s lin-to-circ ratio was close to 0.3. This makes circ*EXOC6B* even harder to detect, because of the expected amount of linear "noise" not contributing to an exact measurement but leading to biased quantifications.

Another striking property makes circ*RMST* a surprising candidate: it is the circRNA in the discovery cohort with the highest expression level across all samples (22.1 junction reads per million), suggesting circ*RMST* is one of the best circRNAs to detect due to low minimum detection sensitivity.

The spatio-temporal regulation of circRNA and its therefore distinct expression patterns make this novel type of (mostly) non-coding RNA an emerging field of biomarker research.

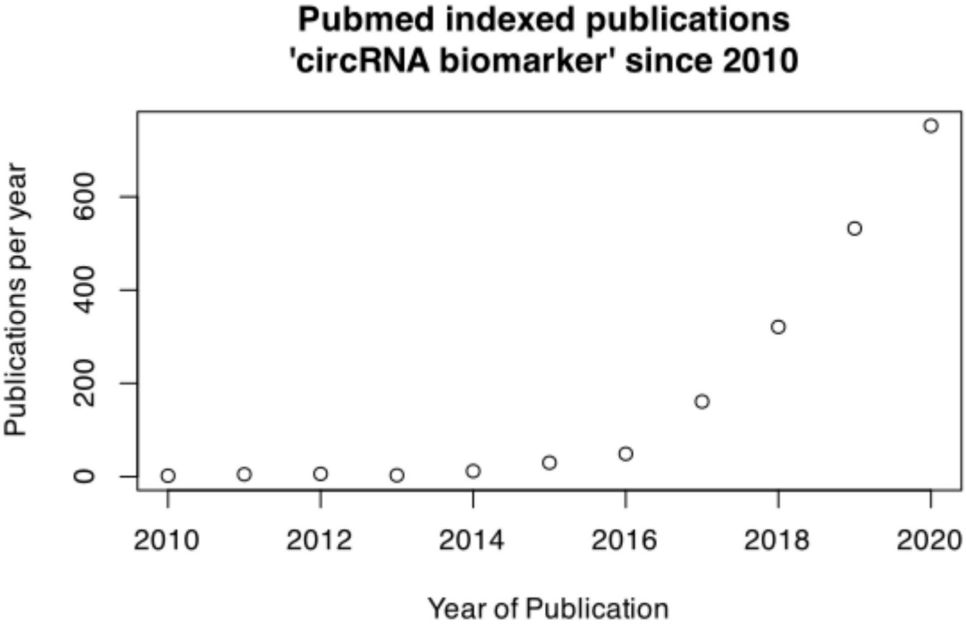


Figure 46: Pubmed indexed publications matching "circRNA biomarker" from 2010-2020 source: *https://pubmed.ncbi.nlm.nih.gov/*

However, there are numerous pitfalls in circRNA-based biomarker research: generally lower agreement between different circRNA detection pipelines indicates a low number of consensus species, and points toward significant false positive rates for each existing algorithm. However, this can be overcome by using multiple approaches at once and only proceeding with the overlapping circRNA species. This, however, can combine the blind spots of each algorithm[102,103].

The *in silico* circRNA research field is based on three main circRNA validation methods: RNAseR treatment of samples and the analysis thereof, called the Circleseq protocol; PCR of the back-splice junction; and northern blot. The shortcomings of the Circleseq protocol are manifold: the sequence depth is a crucial factor in the number of detected circRNAs in a given sample; long circRNAs are RNAseR sensitive[49]; and if not treated long enough, linear RNA can falsify the resulting data. By contrast, PCR is inexpensive and fast, but not every circRNA candidate can be measured with this approach. The length of most circRNAs is limited, hence the design of suitable primers may be challenging. PCR primers also can, if not designed carefully, indicate not only the circular but also the linear isoform, corrupting the whole approach. Other pitfalls are typical for PCR: self-sticking primers that attach to themselves rather than the DNA template[105]; primer sequences that are not unique leading to unspecific amplications; and products that are too large for the PCR reaction, making the desired PCR signal weaker through unfinished synthesis steps.

Northern blotting is also not ideal for circRNA validation. Even though circRNA is more stable than its linear counterpart, it may degrade over time, making the whole analysis time sensitive. This method is also limited in throughput, making it laborious and time consuming for the validation of multiple candidates.

RNA-Seq-based methods, such as those presented here with *circs*, are not immune to mistakes either[49,100.]. RNA-Seq is not an easy, inexpensive or fast process[46]. The sequencing of genetic material using current NGS sequencing machines is still prone to PCR biases and sequencing errors, and can, supplied with material of low quality or quantity, result in poor data quality[43]. Before RNA is sequenced, reverse transcriptase is used to convert the unstable RNA into the more stable DNA molecules that are the ultimate input for all Illumina-based RNA-Seq efforts[43]. This, again can bias the results, and there have been no extensive studies on how selective this conversion step is towards circRNAs.

However, before the emerging and promising biomarkers presented here can be applied, several developments and advancements have to be achieved.

First, more patient-derived samples need to be screened for circ*RMST*, ideally by PCR or RNA-Seq. A high number of samples would be beneficial, ensuring the reliability of this putative biomarker molecule. A potential key finding could be the detection of circ*RMST* in the CSF of WNT MB patients. This would make circRNA screening a faster method of diagnosis, allowing timely risk stratification of the patient before a biosample of the tumor itself can be obtained during resection of the cancerous tissue. Time is crucial in the treatment of cancer, and the ability to identify certain malignant brain cancers through biofluids like CSF would improve the time to diagnosis as well as monitoring disease activity after diagnosis substantially. Exosomal circRNAs have been identified before[92,93] and have even been shown to be enriched in CSF[95], but the ratio of circRNAs in a pediatric brain tumor to circRNAs in its CSF has not yet been identified. CSF samples could be contaminated by blood, indicating another source of noise in this still feasible circRNA biomarker development path. If a circRNA based biomarker for a MB group is found in another malignant brain tumor, such as ependymoma, it could render the candidate unsuitable for further studies. For biomarker development, a high false positive or false negative rate results in an unreliable indicator[139]. Additionally, practical aspects have to be accounted for, such as utility (in the clinic) and practicability (to measure). This refers to the required resources in order to implement the biomarker in the clinical setting; if an expensive machine must be purchased for a single biomarker and new protocols must be established, the biomarker is much harder to use than one with a standard protocol using a new probe. For example, a new set of PCR primers is easy to implement. Lastly, if the biomarker has excellent reliability, but lacks a significant clinical impact (only indicating minor differences that are not actionable), the biomarker will not be implemented.

This thesis demonstrated the remarkable potential of circ*RMST* in sensitivity and specificity, in two independent MB datasets with limited size. Its clinical utility could be immense; early and precise risk stratification can lead to more children surviving this still deadly disease, and lead to less severe side effects in the process. As far as practicability is concerned, the measurement of the circRNA itself is already feasible in almost every standard clinic laboratory equipped with a PCR machine[49,54].

In total, circ*RMST* emerged as a promising WNT MB biomarker, but additional research is required to verify and test its biomarker capacity in the clinical setting.

The performance (measured in RNAseR resistant circRNAs detected in the non-RNAseR treated sample) of the circRNA detection pipeline circs presented here is high (86.8% combining all cell lines / 85.8% for maximal single-cell line), and thus re-using the pipeline can be recommended, especially the snakemake-based circs_snake. However, circRNA analysis typically benefits from strictly trimmed and deeply sequenced paired end data[102], which is recommended as input here as well.

Although the biological function is mostly elusive and remains to be understood, circRNA is still an emerging field that, despite its many pitfalls, holds the promise of being the next frontier in biomarker discovery. Here, we could find a suitable biomarker candidate for WNT MBs, but subsequent studies in this field with a larger number of samples or more sequencing data available could uncover a plethora of suitable circRNAs. This thesis also demonstrated that biomarker discovery can fail at several stages, as seen here in the two other main candidates, circ*EXOC6B* and circ*ISPD*.

Furthermore, this thesis demonstrated that circRNA is a useful additional layer of data, serving as another approach to understand the complex biology of pediatric malignancies. Here, several aspects can lead to further investigations. For example, the circRNA signal holds some informational value, as demonstrated by the MB-group specific circRNA signal, but could not be followed up on in the validation cohort. The *MYC*/RNAseR data set confirmed this trend further, with *MYC* overexpression leading to a generally lower circRNA signal in these cells. In this dataset, as in the validation dataset, this behavior did not reach statistical significance, yet its consistency is striking. This global change of circRNA abundance upon MYC expression perturbations could be caused by a mechanistic role of *MYC* antagonistic to *QKI*, as has been reported previously[75]. However, independent of *MYC*, the general trend of circRNA signal decrease observed here might be of further use in other cancer entities and diseases, as previously demonstrated[91].

6 References

1. Hallmarks of Cancer: The Next Generation I Elsevier Enhanced Reader. doi:10.1016/j.cell.2011.02.013

2. Pollack IF, Agnihotri S, Broniscer A. Childhood brain tumors: current management, biological insights, and future directions: JNSPG 75th Anniversary Invited Review Article. *J Neurosurg Pediatr*. 2019;23(3):261-273. doi:10.3171/2018.10.PEDS18377

3. Barnholtz-Sloan JS, Ostrom QT, Cote D. Epidemiology of Brain Tumors. *Neurol Clin*. 2018;36(3):395-419. doi:10.1016/j.ncl.2018.04.001

4. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol (Berl)*. 2016;131(6):803-820. doi:10.1007/s00401-016-1545-1

5. Arvanitis CD, Ferraro GB, Jain RK. The blood–brain barrier and blood–tumour barrier in brain tumours and metastases. *Nat Rev Cancer*. 2020;20(1):26-41. doi:10.1038/s41568-019-0205-x

6. Haumann R, Videira JC, Kaspers GJL, van Vuurden DG, Hulleman E. Overview of Current Drug Delivery Methods Across the Blood–Brain Barrier for the Treatment of Primary Brain Tumors. *CNS Drugs*. 2020;34(11):1121-1131. doi:10.1007/s40263-020-00766-w

7. Mulhern RK, Merchant TE, Gajjar A, Reddick WE, Kun LE. Late neurocognitive sequelae in survivors of brain tumours in childhood. *Lancet Oncol*. 2004;5(7):399-408. doi:10.1016/S1470-2045(04)01507-4

8. Olesen J, Gustavsson A, Svensson M, et al. The economic cost of brain disorders in Europe: Economic cost of brain disorders in Europe. *Eur J Neurol*. 2012;19(1):155-162. doi:10.1111/j.1468-1331.2011.03590.x

9. Northcott PA, Robinson GW, Kratz CP, et al. Medulloblastoma. *Nat Rev Dis Primer*. 2019;5(1). doi:10.1038/s41572-019-0063-6

10. Guerreiro Stucklin AS, Ramaswamy V, Daniels C, Taylor MD. Review of molecular classification and treatment implications of pediatric brain tumors: *Curr Opin Pediatr*. 2018;30(1):3-9. doi:10.1097/MOP.0000000000000562

11. Othman RT, Kimishi I, Bradshaw TD, et al. Overcoming multiple drug resistance mechanisms in medulloblastoma. *Acta Neuropathol Commun*. 2014;2(1):57. doi:10.1186/2051-5960-2-57

12. Kralik SF, Ho CY, Finke W, Buchsbaum JC, Haskins CP, Shih C-S. Radiation Necrosis in Pediatric Patients with Brain Tumors Treated with Proton Radiotherapy. *Am J Neuroradiol*. 2015;36(8):1572-1578. doi:10.3174/ajnr.A4333

13. Chapter 8 - Cerebellum: Development and Medulloblastoma I Elsevier Enhanced Reader. doi:10.1016/B978-0-12-380916-2.00008-5

14. Kool M, Koster J, Bunt J, et al. Integrated Genomics Identifies Five Medulloblastoma Subtypes with Distinct Genetic Profiles, Pathway Signatures and Clinicopathological Features. Hide W, ed. *PLoS ONE*. 2008;3(8):e3088. doi:10.1371/journal.pone.0003088

15. Remke M, Hielscher T, Northcott PA, et al. Adult Medulloblastoma Comprises Three Major Molecular Variants. *J Clin Oncol*. 2011;29(19):2717-2723. doi:10.1200/JCO.2011.34.9373

16. Remke M, Ramaswamy V, Taylor MD. Medulloblastoma molecular dissection: the way toward targeted therapy. *Curr Opin Oncol*. 2013;25(6):674-681. doi:10.1097/CCO.0000000000000008

17. Archer TC, Ehrenberger T, Mundt F, et al. Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell*. 2018;34(3):396-410.e8. doi:10.1016/j.ccell.2018.08.004

18. Cavalli FMG, Remke M, Rampasek L, et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell*. 2017;31(6):737-754.e6. doi:10.1016/j.ccell.2017.05.005

19. Remke M, Ramaswamy V. Infant medulloblastoma — learning new lessons from old strata. *Nat Rev Clin Oncol*. 2018;15(11):659-660. doi:10.1038/s41571-018-0071-6

20. Ellison DW, Kocak M, Dalton J, et al. Definition of Disease-Risk Stratification Groups in Childhood Medulloblastoma Using Combined Clinical, Pathologic, and Molecular Variables. *J Clin Oncol*. 2011;29(11):1400-1407. doi:10.1200/JCO.2010.30.2810

21. Mulhern RK, Palmer SL, Merchant TE, et al. Neurocognitive Consequences of Risk-Adapted Therapy for Childhood Medulloblastoma. *J Clin Oncol*. 2005;23(24):5511-5519. doi:10.1200/JCO.2005.00.703

22. Manoranjan B, Venugopal C, Bakhshinyan D, et al. Wnt activation as a therapeutic strategy in medulloblastoma. *Nat Commun*. 2020;11(1):4323. doi:10.1038/s41467-020-17953-4

23. Juraschka K, Taylor MD. Medulloblastoma in the age of molecular subgroups: a review. *J Neurosurg Pediatr*. 2019;24(4):353-363. doi:10.3171/2019.5.PEDS18381

24. Hovestadt V, Smith KS, Bihannic L, et al. Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature*. 2019;572(7767):74-79. doi:10.1038/s41586-019-1434-6

25. Taylor MD, Northcott PA, Korshunov A, et al. Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol (Berl)*. 2012;123(4):465-472. doi:10.1007/s00401-011-0922-z

26. Rathi KS, Arif S, Koptyra M, et al. A transcriptome-based classifier to determine molecular subtypes in medulloblastoma. *PLOS Comput Biol*.:15.

27. Gendoo DMA, Smirnov P, Lupien M, Haibe-Kains B. Personalized diagnosis of medulloblastoma subtypes across patients and model systems. Genomics. 2015 Aug;106(2):96-106. doi: 10.1016/j.ygeno.2015.05.002.

28. Schwalbe EC. Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. Lancet Oncol. 2017 Jul;18(7):958-971. doi: 10.1016/S1470-2045(17)30243-7. 2017;18:14.

29. Remke M, Hielscher T, Korshunov A, et al. *FSTL5* Is a Marker of Poor Prognosis in Non-WNT/Non-SHH Medulloblastoma. *J Clin Oncol*. 2011;29(29):3852-3861. doi:10.1200/JCO.2011.36.2798

30. Massimino M, Biassoni V, Gandola L. Childhood medulloblastoma. Crit Rev Oncol Hematol. 2016 ;105:35-51. doi: 10.1016/j.critrevonc.2016.05.012.

31. Zhukova N, Ramaswamy V, Remke M, et al. Subgroup-Specific Prognostic Implications of *TP53* Mutation in Medulloblastoma. *J Clin Oncol*. 2013;31(23):2927-2935. doi:10.1200/JCO.2012.48.5052

32. Khanna V, Achey RL, Ostrom QT, et al. Incidence and survival trends for medulloblastomas in the United States from 2001 to 2013. J Neurooncol. 2017;135(3):433-441. doi:10.1007/s11060-017-2594-6

33. Forget A, Martignetti L, Puget S, et al. Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling. *Cancer Cell*. 2018;34(3):379-395.e7. doi:10.1016/j.ccell.2018.08.002

34. Hovestadt V, Remke M, Kool M, et al. Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol (Berl)*. 2013;125(6):913-916. doi:10.1007/s00401-013-1126-5

35. Ramaswamy V, Samuel N, Remke M. Can miRNA-based real-time PCR be used to classify medulloblastomas? *CNS Oncol*. 2014;3(3):173-175. doi:10.2217/cns.14.14

36. Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555(7697):469-474. doi:10.1038/nature26000

37. Ramaswamy V, Remke M, Bouffet E, et al. Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. *Acta Neuropathol (Berl)*. 2016;131(6):821-831. doi:10.1007/s00401-016-1569-6

38. Eberhart CG, Tihan T, Burger PC. Nuclear localization and mutation of beta-catenin in medulloblastomas. J Neuropathol Exp Neurol. 2000 ;59(4):333-7. doi: 10.1093/jnen/59.4.333.

39. Orr BA, Clay MR, Pinto EM, Kesserwan C. An update on the central nervous system manifestations of Li–Fraumeni syndrome. *Acta Neuropathol (Berl)*. 2020;139(4):669-687. doi:10.1007/s00401-019-02055-3

40. Polkinghorn WR, Tarbell NJ. Medulloblastoma: tumorigenesis, current clinical paradigm, and efforts to improve risk stratification. *Nat Clin Pract Oncol*. 2007;4(5):295-304. doi:10.1038/ncponc0794

41. Northcott PA, Shih DJH, Peacock J, et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*. 2012;488(7409):49-56. doi:10.1038/nature11327

42. Kumari A, Folk W, Sakamuro D. The Dual Roles of MYC in Genomic Instability and Cancer Chemoresistance. *Genes*. 2017;8(6):158. doi:10.3390/genes8060158

43. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484

44. Yang H-P. Reference Based RNA-Seq Data Analysis. Available at https://bioinformatics.uconn.edu/reference-based-rna-seq-data-analysis/, 21. April 2021 :66.

45. Scholes AN, Lewis JA. Comparison of RNA isolation methods on RNA-Seq: implications for differential expression and meta-analyses. *BMC Genomics*.

2020;21(1):249. doi:10.1186/s12864-020-6673-2

46. HiSeq 2500 System Guide. Available at https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/hiseq2500/hiseq-2500-system-guide-15035786-03.pdf, 21. April 2021:96.

47. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods*. 2017;14(2):135-139. doi:10.1038/nmeth.4106

48. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

49. Szabo L, Salzman J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nat Rev Genet*. 2016;17(11):679-692. doi:10.1038/nrg.2016.114

50. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11(1):94. doi:10.1186/1471-2105-11-94

51. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*. 2011;8(6):469-477. doi:10.1038/nmeth.1613

52. Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. Cell-Type Specific Features of Circular RNA Expression. Moran JV, ed. *PLoS Genet*. 2013;9(9):e1003777. doi:10.1371/journal.pgen.1003777

53. Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*. 2013;495(7441):333-338. doi:10.1038/nature11928

54. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. Preiss T, ed. *PLoS ONE*. 2012;7(2):e30733. doi:10.1371/journal.pone.0030733

55. Wilusz JE. Circular RNAs: Unexpected outputs of many protein-coding genes. *RNA Biol*. 2017;14(8):1007-1017. doi:10.1080/15476286.2016.1227905

56. Jakobi T, Dieterich C. Computational approaches for circular RNA analysis. *Wiley Interdiscip Rev RNA*. 2019;10(3):e1528. doi:10.1002/wrna.1528

57. Rybak-Wolf A, Stottmeister C, Glažar P, et al. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol Cell*. 2015;58(5):870-885. doi:10.1016/j.molcel.2015.03.027

58. Dube U, Del-Aguila JL, Li Z, et al. An atlas of cortical circular RNA expression in Alzheimer disease brains demonstrates clinical and pathological associations. Nat Neurosci. 2019;22(11):1903-1912. doi:10.1038/s41593-019-0501-5

59. Gokool A, Anwar F, Voineagu I. The Landscape of Circular RNA Expression in the Human Brain. *Biol Psychiatry*. 2020;87(3):294-304. doi:10.1016/j.biopsych.2019.07.029

60. Hanan M, Soreq H, Kadener S. CircRNAs in the brain. *RNA Biol*. 2017;14(8):1028-1034. doi:10.1080/15476286.2016.1255398

61. Venø MT, Hansen TB, Venø ST, et al. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. *Genome Biol*. 2015;16(1). doi:10.1186/s13059-015-0801-3

62. Wang PL, Bao Y, Yee M-C, et al. Circular RNA Is Expressed across the Eukaryotic Tree of Life. Preiss T, ed. *PLoS ONE*. 2014;9(3):e90859. doi:10.1371/journal.pone.0090859

63. Demongeot J, Seligmann H. Spontaneous evolution of circular codes in theoretical minimal RNA rings. *Gene*. 2019;705:95-102. doi:10.1016/j.gene.2019.03.069

64. Hassanin, A. (2020). The SARS-CoV-2-like virus found in captive pangolins from Guangdong should be better sequenced. BioRxiv 2021.

65. Toptan T, Abere B, Nalesnik MA, et al. Circular DNA tumor viruses make circular RNAs. *Proc Natl Acad Sci*. 2018;115(37):E8737-E8745. doi:10.1073/pnas.1811728115

66. Ungerleider NA, Jain V, Wang Y, et al. Comparative Analysis of Gammaherpesvirus Circular RNA Repertoires: Conserved and Unique Viral Circular RNAs. Longnecker RM, ed. *J Virol*. 2018;93(6):e01952-18, /jvi/93/6/JVI.01952-18.atom. doi:10.1128/JVI.01952-18

67. Huang JT, Chen JN, Gong LP, et al. Identification of virus-encoded circular RNA. Virology. 2019;529:144-151. doi:10.1016/j.virol.2019.01.014

68. Salzman J. Circular RNA Expression: Its Potential Regulation and Function. *Trends Genet*. 2016;32(5):309-316. doi:10.1016/j.tig.2016.03.002

69. Robic A, Demars J, Kühn C. In-Depth Analysis Reveals Production of Circular RNAs from Non-Coding Sequences. *Cells*. 2020;9(8):1806. doi:10.3390/cells9081806

70. Panda AC, De S, Grammatikakis I, et al. High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs. *Nucleic Acids Res*. 2017;45(12):e116-e116. doi:10.1093/nar/gkx297

71. Stagsted LV, Nielsen KM, Daugaard I, Hansen TB. Noncoding AUG circRNAs constitute an abundant and conserved subclass of circles. *Life Sci Alliance*. 2019;2(3):e201900398. doi:10.26508/lsa.201900398

72. Jeck WR, Sorrentino JA, Wang K, et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*. 2013;19(2):141-157. doi:10.1261/rna.035667.112

73. Kristensen LS, Hansen TB, Venø MT, Kjems J. Circular RNAs in cancer: opportunities and challenges in the field. *Oncogene*. 2018;37(5):555-565. doi:10.1038/onc.2017.361

74. Stagsted LVW. Title: The RNA-binding protein SFPQ preserves long-intron splicing and regulates circRNA biogenesis. :33.

75. Conn SJ, Pillman KA, Toubia J, et al. The RNA Binding Protein Quaking Regulates Formation of circRNAs. *Cell*. 2015;160(6):1125-1134. doi:10.1016/j.cell.2015.02.014

76. Knupp D, Cooper DA, Saito Y, Darnell RB, Miura P. NOVA2 regulates neural circRNA biogenesis. *bioRxiv* 2021:2021.05.02.442201. doi:10.1101/2021.05.02.442201

77. Starke S, Jost I, Rossbach O, et al. Exon Circularization Requires Canonical Splice Signals. *Cell Rep*. 2015;10(1):103-111. doi:10.1016/j.celrep.2014.12.002

78. Wilusz JE. A 360° view of circular RNAs: From biogenesis to functions. *Wiley Interdiscip Rev RNA*. 2018;9(4):e1478. doi:10.1002/wrna.1478

79. Zhang X-O, Dong R, Zhang Y, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res*. 2016;26(9):1277-1287. doi:10.1101/gr.202895.115

80. Liang D, Tatomer DC, Luo Z, et al. The Output of Protein-Coding Genes Shifts to Circular RNAs When the Pre-mRNA Processing Machinery Is Limiting. *Mol Cell*. 2017;68(5):940-954.e3. doi:10.1016/j.molcel.2017.10.034

81. Hansen TB, Jensen TI, Clausen BH, et al. Natural RNA circles function as efficient microRNA sponges. *Nature*. 2013;495(7441):384-388. doi:10.1038/nature11993

82. Hao Z, Hu S, Liu Z, Song W, Zhao Y, Li M. Circular RNAs: Functions and Prospects in Glioma. J Mol Neurosci. 2019;67(1):72-81. doi:10.1007/s12031-018-1211-2

83. Pamudurti NR, Bartok O, Jens M, et al. Translation of CircRNAs. *Mol Cell*. 2017;66(1):9-21.e7. doi:10.1016/j.molcel.2017.02.021

84. Chekulaeva M, Rajewsky N. Roles of Long Noncoding RNAs and Circular RNAs in Translation. *Cold Spring Harb Perspect Biol*. 2019;11(6):a032680. doi:10.1101/cshperspect.a032680

85. Li HM, Ma XL, Li HG. Intriguing circles: Conflicts and controversies in circular RNA research. Wiley Interdiscip Rev RNA. 2019;10(5):e1538. doi:10.1002/wrna.1538

86. Arnaiz E, Sole C, Manterola L, Iparraguirre L, Otaegui D, Lawrie CH. CircRNAs and cancer: Biomarkers and master regulators. Semin Cancer Biol. 2019;58:90-99. doi:10.1016/j.semcancer.2018.12.002

87. Brown JR, Chinnaiyan AM. The Potential of Circular RNAs as Cancer Biomarkers. Cancer Epidemiol Biomarkers Prev. 2020;29(12):2541-2555. doi:10.1158/1055-9965.EPI-20-0796

88. Vo JN, Cieslik M, Zhang Y, et al. The Landscape of Circular RNA in Cancer. *Cell*. 2019;176(4):869-881.e13. doi:10.1016/j.cell.2018.12.021

89. Bonizzato A, Gaffo E, te Kronnie G, Bortoluzzi S. CircRNAs in hematopoiesis and hematological malignancies. *Blood Cancer J*. 2016;6(10):e483-e483. doi:10.1038/bcj.2016.81

90. Lv T, Miao Y-F, Jin K, et al. Dysregulated circular RNAs in medulloblastoma regulate proliferation and growth of tumor cells via host genes. *Cancer Med*. 2018;7(12):6147-6157. doi:10.1002/cam4.1613

91. Okholm TLH, Nielsen MM, Hamilton MP, et al. Circular RNA expression is abundant and correlated to aggressiveness in early-stage bladder cancer. *Npj Genomic Med*. 2017;2(1). doi:10.1038/s41525-017-0038-z

92. Li Y, Zheng Q, Bao C, et al. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res*. 2015;25(8):981-984. doi:10.1038/cr.2015.82

93. Wang Y, Liu J, Ma J, et al. Exosomal circRNAs: biogenesis, effect and application in human diseases. *Mol Cancer*. 2019;18(1). doi:10.1186/s12943-019-1041-z

94. Seimiya T, Otsuka M, Iwata T, et al. Emerging Roles of Exosomal Circular RNAs in

Cancer. *Front Cell Dev Biol*. 2020;8:568366. doi:10.3389/fcell.2020.568366

95. Hulstaert E, Morlion A, Avila Cobos F, et al. Charting Extracellular Transcriptomes in The Human Biofluid RNA Atlas. *Cell Rep*. 2020;33(13):108552. doi:10.1016/j.celrep.2020.108552

96. Chen L, Wang C, Sun H, et al. The bioinformatics toolbox for circRNA discovery and analysis. Brief Bioinform. 2021;22(2):1706-1728. doi:10.1093/bib/bbaa001

97. Dahl M, Daugaard I, Andersen MS, et al. Enzyme-free digital counting of endogenous circular RNA molecules in B-cell malignancies. *Lab Invest*. 2018;98(12):1657-1669. doi:10.1038/s41374-018-0108-6

98. Ma XK, Wang MR, Liu CX, et al. CIRCexplorer3: A CLEAR Pipeline for Direct Comparison of Circular and Linear RNA Expression. Genomics Proteomics Bioinformatics. 2019;17(5):511-521. doi:10.1016/j.gpb.2019.11.004

99. Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, Yang L. Complementary Sequence-Mediated Exon Circularization. *Cell*. 2014;159(1):134-147. doi:10.1016/j.cell.2014.09.001

100. Hansen TB, Venø MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. *Nucleic Acids Res*. 2016;44(6):e58-e58. doi:10.1093/nar/gkv1458

101. Barrett SP, Salzman J. Circular RNAs: analysis, expression and potential functions. *Development*. 2016;143(11):1838-1847. doi:10.1242/dev.128074

102. Hansen TB. Improved circRNA Identification by Combining Prediction Algorithms. *Front Cell Dev Biol*. 2018;6. doi:10.3389/fcell.2018.00020

103. Gaffo E, Bonizzato A, Kronnie G, Bortoluzzi S. CirComPara: A Multi-Method Comparative Bioinformatics Pipeline to Detect and Study circRNAs from RNA-seq Data. *Non-Coding RNA*. 2017;3(1):8. doi:10.3390/ncrna3010008

104. Ahmadov U, Bendikas MM, Ebbesen KK, et al. Distinct circular RNA expression profiles in pediatric ependymomas. Brain Pathol. 2021;31(2):387-392. doi:10.1111/bpa.12922

105. Jakobi T, Uvarovskii A, Dieterich C. circtools-a one-stop software solution for circular RNA research. Bioinformatics. 2019;35(13):2326-2328. doi:10.1093/bioinformatics/bty948

106. Boss M, Arenz C. A Fast and Easy Method for Specific Detection of Circular RNA by Rolling-Circle Amplification. *ChemBioChem*. 2020;21(6):793-796. doi:10.1002/cbic.201900514

107. Zhu J, Ye J, Zhang L, et al. Differential Expression of Circular RNAs in Glioblastoma Multiforme and Its Correlation with Prognosis. *Transl Oncol*. 2017;10(2):271-279. doi:10.1016/j.tranon.2016.12.006

108. Liu M, Wang Q, Shen J, Yang BB, Ding X. Circbank: a comprehensive database for circRNA with standard nomenclature. *RNA Biol*. 2019;16(7):899-905. doi:10.1080/15476286.2019.1600395

109. Vromman M, Vandesompele J, Volders PJ. Closing the circle: current state and perspectives of circular RNA databases. Brief Bioinform. 2021;22(1):288-297. doi:10.1093/bib/bbz175

110. Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. *RNA*. 2014;20(11):1666-1670. doi:10.1261/rna.043687.113

111. Lyu Y, Caudron-Herger M, Diederichs S. circ2GO: A Database Linking Circular RNAs to Gene Function. *Cancers*. 2020;12(10):2975. doi:10.3390/cancers12102975

112. Metge F, Czaja-Hasse LF, Reinhardt R, Dieterich C. FUCHS—towards full circular RNA characterization using RNAseq. *PeerJ*. 2017;5:e2934. doi:10.7717/peerj.2934

113. Chen L, Wang F, Bruggeman EC, Li C, Yao B. circMeta: a unified computational framework for genomic feature annotation and differential expression analysis of circular RNAs. Bioinformatics. 2020;36(2):539-545. doi:10.1093/bioinformatics/btz606

114. Aufiero S, Reckman YJ, Tijsen AJ, Pinto YM, Creemers EE. circRNAprofiler: an R-based computational framework for the downstream analysis of circular RNAs. *BMC Bioinformatics*. 2020;21(1):164. doi:10.1186/s12859-020-3500-3

115. Ghosal S, Das S, Sen R, Basak P, Chakrabarti J. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. Front Genet. 2013;4:283. Published 2013 Dec 10. doi:10.3389/fgene.2013.00283

116. Ungerleider N, Flemington E. SpliceV: analysis and publication quality printing of linear and circular RNA splicing, expression and regulation. *BMC Bioinformatics*. 2019;20(1). doi:10.1186/s12859-019-2865-7

117. Feng J, Xiang Y, Xia S, et al. CircView: a visualization and exploration tool for circular RNAs. *Brief Bioinform*. 2019;20(3):745-751. doi:10.1093/bib/bbx070

118. Ferrero G, Licheri N, Coscujuela Tarrero L, et al. Docker4Circ: A Framework for the Reproducible Characterization of circRNAs from RNA-Seq Data. *Int J Mol Sci*. 2019;21(1):293. doi:10.3390/ijms21010293

119. Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics*. 2016;32(7):1094-1096. doi:10.1093/bioinformatics/btv656

120. Akers NK, Schadt EE, Losic B. STAR Chimeric Post for rapid detection of circular RNA and fusion transcripts. Valencia A, ed. *Bioinformatics*. 2018;34(14):2364-2370. doi:10.1093/bioinformatics/bty091

121. Chaabane M, Williams RM, Stephens AT, Park JW. circDeep: deep learning approach for circular RNA classification from other long non-coding RNA. Bioinformatics. 2020;36(1):73-80. doi:10.1093/bioinformatics/btz537

122. Song X, Zhang N, Han P, et al. Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res*. 2016;44(9):e87-e87. doi:10.1093/nar/gkw075

123. Wang J, Wang L. Deep learning of the back-splicing code for circular RNA formation. Bioinformatics. 2019;35(24):5235-5242. doi:10.1093/bioinformatics/btz382

124. Yoshimoto R, Rahimi K, Hansen TB, Kjems J, Mayeda A. Biosynthesis of Circular RNA ciRS-7/CDR1as Is Mediated by Mammalian-wide Interspersed Repeats. *iScience*. 2020;23(7):101345. doi:10.1016/j.isci.2020.101345

125. Mölder F, Jablonski KP, Letcher B, et al. Sustainable data analysis with Snakemake. *F1000Research*. 2021;10:33. doi:10.12688/f1000research.29032.1

126. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333-337. doi:10.1038/nmeth.2810

127. Ng S-Y, Bogu GK, Soh BS, Stanton LW. The Long Noncoding RNA RMST Interacts with SOX2 to Regulate Neurogenesis. *Mol Cell*. 2013;51(3):349-359. doi:10.1016/j.molcel.2013.07.017

128. Pajtler KW, Witt H, Sill M, et al. Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell*. 2015;27(5):728-743. doi:10.1016/j.ccell.2015.04.002

129. Wang L, Liu D, Wu X, et al. Long non-coding RNA (LncRNA) RMST in triple-negative breast cancer (TNBC): Expression analysis and biological roles research. *J Cell Physiol*. 2018;233(10):6603-6612. doi:https://doi.org/10.1002/jcp.26311

130. Uhde CW, Vives J, Jaeger I, Li M. Rmst Is a Novel Marker for the Mouse Ventral Mesencephalic Floor Plate and the Anterior Dorsal Midline Cells. Riley B, ed. *PLoS ONE*. 2010;5(1):e8641. doi:10.1371/journal.pone.0008641

131. Izuogu OG, Alhasan AA, Mellough C, et al. Analysis of human ES cell differentiation establishes that the dominant isoforms of the lncRNAs RMST and FIRRE are circular. *BMC Genomics*. 2018;19(1). doi:10.1186/s12864-018-4660-7

132. Yu C-Y, Kuo H-C. The Trans-Spliced Long Noncoding RNA tsRMST Impedes Human Embryonic Stem Cell Differentiation Through WNT5A-Mediated Inhibition of the Epithelial-to-Mesenchymal Transition. *Stem Cells Dayt Ohio*. 2016;34(8):2052-2062. doi:10.1002/stem.2386

133. Cirak S, Foley AR, Herrmann R, et al. ISPD gene mutations are a common cause of congenital and limb-girdle muscular dystrophies. *Brain*. 2013;136(1):269-281. doi:10.1093/brain/aws312

134. Magri F, Colombo I, Del Bo R, et al. ISPD mutations account for a small proportion of Italian Limb Girdle Muscular Dystrophy cases. *BMC Neurol*. 2015;15(1):172. doi:10.1186/s12883-015-0428-8

135. Mahmoudi E, Kiltschewskij D, Fitzsimmons C, Cairns MJ. Depolarization-Associated CircRNA Regulate Neural Gene Expression and in Some Cases May Function as Templates for Translation. *Cells*. 2019;9(1):25. doi:10.3390/cells9010025

136. Evers C, Maas B, Koch KA, et al. Mosaic deletion of EXOC6B: Further evidence for an important role of the exocyst complex in the pathogenesis of intellectual disability. *Am J Med Genet A*. 2014;164(12):3088-3094. doi:https://doi.org/10.1002/ajmg.a.36770

137. Girisha KM, Kortüm F, Shah H, et al. A novel multiple joint dislocation syndrome associated with a homozygous nonsense variant in the EXOC6B gene. Eur J Hum Genet. 2016;24(8):1206-1210. doi:10.1038/ejhg.2015.261

138. Frühmesser A, Blake J, Haberlandt E, et al. Disruption of EXOC6B in a patient with developmental delay, epilepsy, and a de novo balanced t(2;8) translocation. Eur J Hum Genet. 2013;21(10):1177-1180. doi:10.1038/ejhg.2013.18

139. Strimbu K, Tavel JA. What are biomarkers?. Curr Opin HIV AIDS. 2010;5(6):463-466. doi:10.1097/COH.0b013e32833ed177

7 Appendix

7.1 Supplementary figures

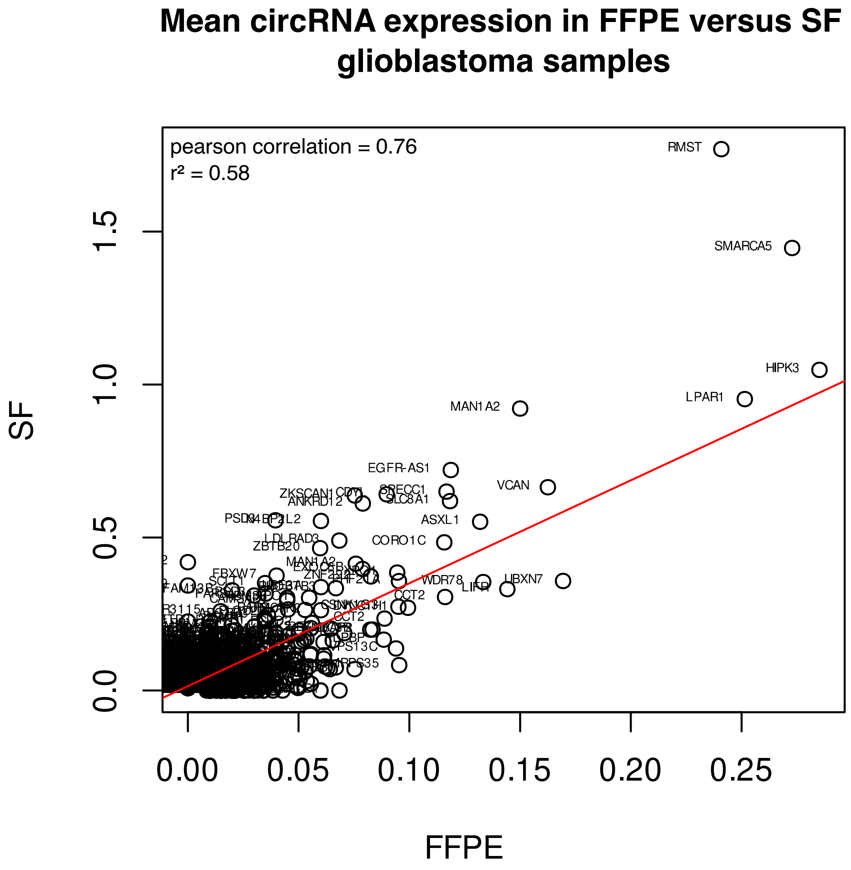


Figure 47: Relative signal loss in matched formalin fixated paraffin embedded (FFPE) and snap frozen (SF) glioblastoma samples. Red linear trend line. Gene names correspond to parental gene annotation.

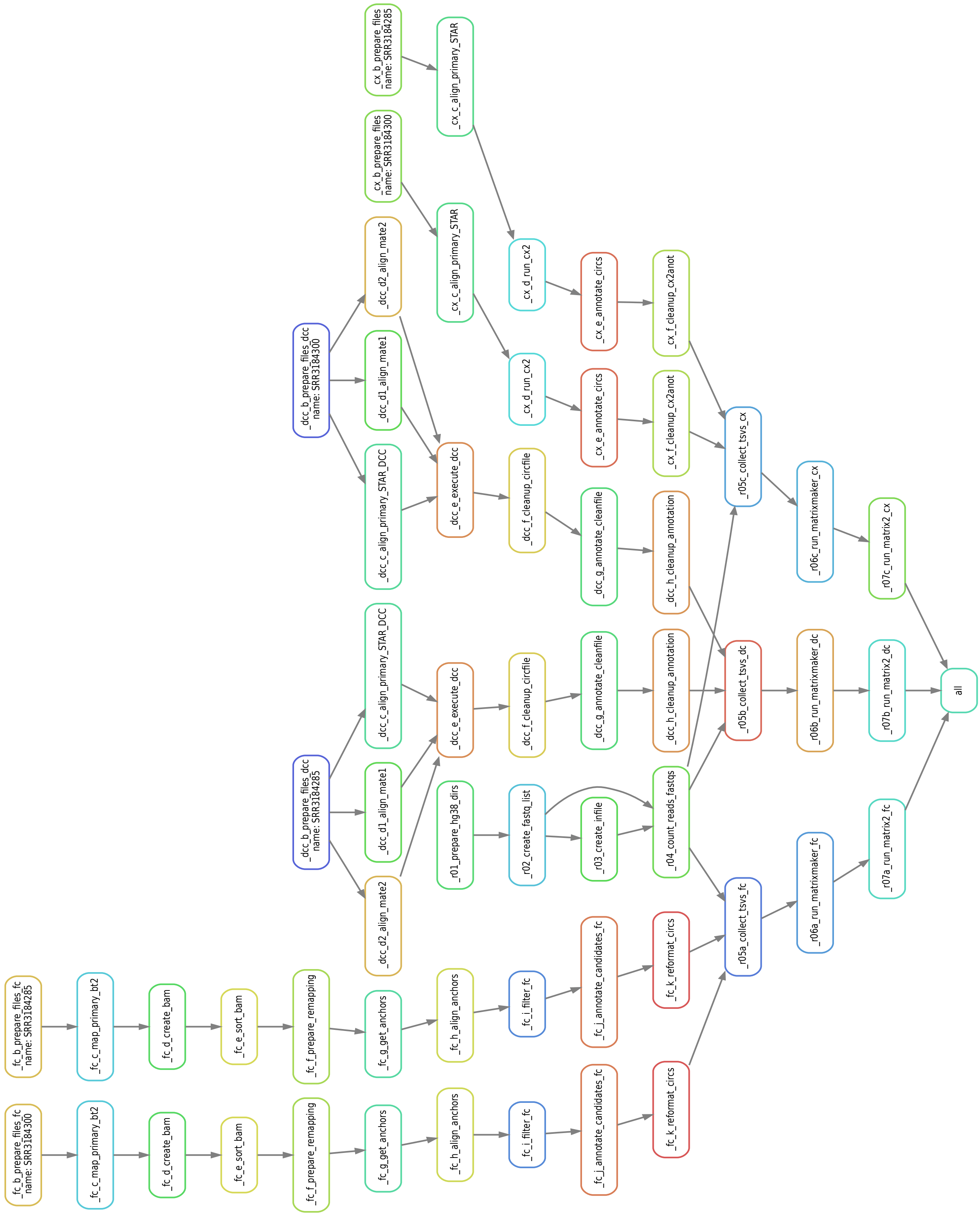


Figure 48: DAG resulting from circs_snake on 2 samples of paired-end data as of version 10.05.2021. Rule “all” does only collect output files.

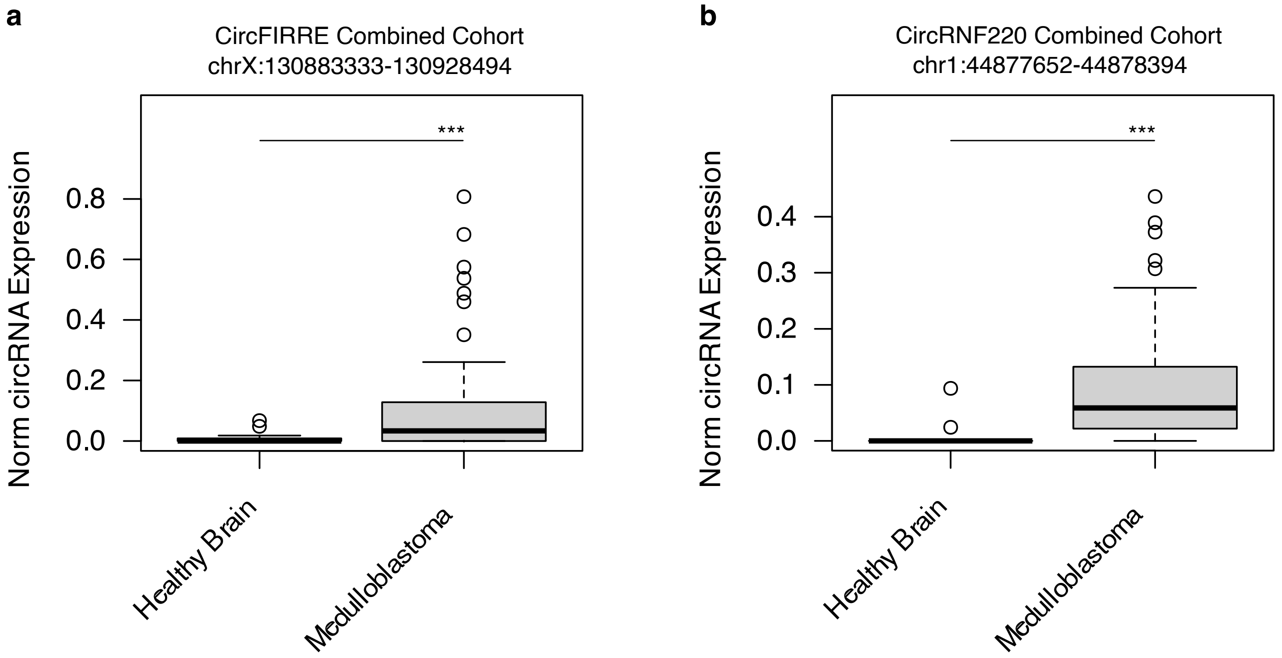


Figure 49: Circular RNA biomarkers for medulloblastoma compared to healthy fetal brain tissue. Medulloblastoma cohort combines discovery and validation cohorts. ***=p-value <0.001.

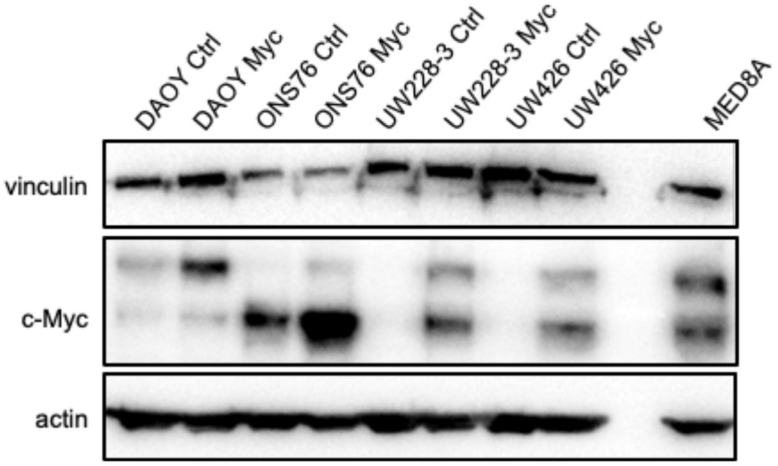


Figure 50: Western blot of c-MYC in *MYC* RNAseR samples.
Myc= *MYC* overexpression samples, Ctrl = control samples.

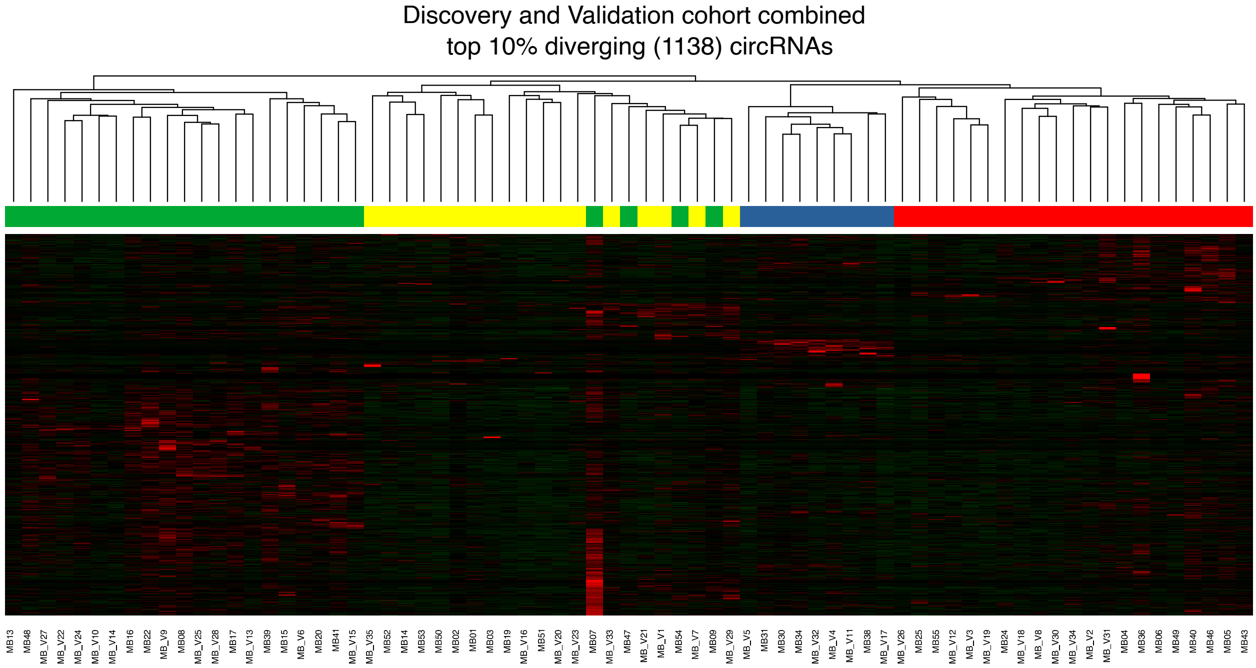


Figure 51: Heatmap of the combined discovery and validation cohort. Shown are the top 10% differentially expressed circular RNAs. Dendrogram is based on Pearson's dissimilarity. colors based on circRNA MB group. blue = WNT MB, red = SHH MB, yellow= Group 3 MB, green = Group 4 MB.

7.2 Supplementary tables

The input data from the *MYC*/RNAseR data set is shown below in Table 9.

Table 9: Table of RNA-Seq preparation metrics for the *MYC*/RNAseR dataset used in this thesis.

Sample Name	RNA isolation method	determination of concentration	Conc (ng/µl)	volume (~µl)	treatment
DAOY_GFP-*MYC*	Maxwell	NanoDrop	640,2	15	Total RNA
DAOY_GFP ctrl.	Maxwell	NanoDrop	708,6	15	Total RNA
ONS76_GFP-*MYC*	Maxwell	NanoDrop	1237,6	15	Total RNA
ONS76_GFP ctrl.	Maxwell	NanoDrop	1342,9	15	Total RNA
UW228-3_GFP-*MYC*	Maxwell	NanoDrop	795,3	15	Total RNA
UW228-3_GFP ctrl.	Maxwell	NanoDrop	800	15	Total RNA
DAOY_GFP-*MYC*	Maxwell	NanoDrop	54,2	~29	RnaseR treated
DAOY_GFP ctrl.	Maxwell	NanoDrop	55,1	~29	RnaseR treated
ONS76_GFP-*MYC*	Maxwell	NanoDrop	53,8	~29	RnaseR treated
ONS76_GFP ctrl.	Maxwell	NanoDrop	49,4	~29	RnaseR treated
UW228-3_GFP-*MYC*	Maxwell	NanoDrop	65,8	~29	RnaseR treated
UW228-3_GFP ctrl.	Maxwell	NanoDrop	64,1	~29	RnaseR treated

Table 10: RNA-Seq MYC quantifications across MYC/RNAseR dataset.

Sample name	Original name	Status	Cell line	Treatment	MYC linear RNA reads
RICK_01_S3	DAOY_GFP-MYC	MYC	DAOY	Total RNA	974
RICK_02_S4	DAOY_GFP ctrl.	Control	DAOY	Total RNA	154
RICK_03_S1	ONS76_GFP-MYC	MYC	ONS76	Total RNA	2924
RICK_04_S2	ONS76_GFP ctrl.	Control	ONS76	Total RNA	2591
RICK_05_S1	UW228-3_GFP-MYC	MYC	UW228-3	Total RNA	4277
RICK_06_S2	UW228-3_GFP ctrl.	Control	UW228-3	Total RNA	500
RICK_07_S3	DAOY_GFP-MYC	MYC	DAOY	RnaseR treated	373
RICK_08_S4	DAOY_GFP ctrl.	Control	DAOY	RnaseR treated	20
RICK_09_S1	ONS76_GFP-MYC	MYC	ONS76	RnaseR treated	940
RICK_10_S2	ONS76_GFP ctrl.	Control	ONS76	RnaseR treated	259
RICK_11_S3	UW228-3_GFP-MYC	MYC	UW228-3	RnaseR treated	2067
RICK_12_S4	UW228-3_GFP ctrl.	Control	UW228-3	RnaseR treated	73

Table 11: Pan-cohort circRNA biomarker species for circ medulloblastoma RNA-based WNT medulloblastoma group. Circular RNA information according to DCC.

coordinates	strand	gene	refseqid	circ_id_circbase
chr12:97886238-97954825	+	RMST	NR_024037	unknown
chr17:63739185-63746842	+	CEP112	NM_001037325	hsa_circ_0002910
chr11:92085261-92088570	-	FAT3	NM_001008781	hsa_circ_0000348
chr11:36248634-36248980	-	LDLRAD3	NM_174902	hsa_circ_0006988

chr12:97886238-97924637	+	RMST	NR_024037	hsa_circ_0027821
chr9:115030328-115060196	+	MIR3134	NR_036085	hsa_circ_0003500
chr1:62321701-62350080	-	PATJ	NM_176877	hsa_circ_0012779
chr9:115024714-115060196	-	MIR3134	NR_036085	hsa_circ_0008192
chr9:73442762-73479427	+	TRPM3	NM_020952	unknown
chr2:223084858-223097002	+	PAX3	NM_001127366	hsa_circ_0007333
chr3:89390065-89391240	-	EPHA3	NM_182644	hsa_circ_0066598
chr3:89456418-89480509	-	EPHA3	NM_005233	unknown
chr3:89456418-89499520	-	EPHA3	NM_005233	hsa_circ_0066601
chr4:42505466-42546003	+	ATP8A1	NM_001105529	hsa_circ_0069613
chr6:94066434-94068129	+	EPHA7	NM_004440	hsa_circ_0077398
chr9:115013208-115060196	+	MIR3134	NR_036085	hsa_circ_0003458
chr9:114842353-114875148	+	MIR3134	NR_036085	unknown
chr9:73477823-73479427	+	TRPM3	NM_020952	unknown
chr3:107910367-107932868	-	IFT57	NM_018010	hsa_circ_0066741
chr22:29517344-29521404	-	KREMEN1	NM_001039570	hsa_circ_0004547
chr4:183522076-183550042	+	TENM3	NM_001080477	hsa_circ_0071480
chr2:159165944-159201830	+	CCDC148	NM_138803	hsa_circ_0056768
chr4:108603170-108622441	+	PAPSS1	NM_005443	hsa_circ_0006935
chr5:145144493-145205763	+	PRELID2	NM_182960	hsa_circ_0008132
chr12:128899276-128900165	-	TMEM132C	NM_001136103	unknown
chr2:107446521-107460490	+	ST6GAL2	NM_001142351	hsa_circ_0055954
chr1:210186977-210194599	-	SYT14	NM_001146261	hsa_circ_0016334
chr9:73376516-73399195	+	TRPM3	NM_020952	unknown
chr4:42618049-42629126	-	ATP8A1	NM_001105529	unknown
chr2:178681555-178705110	+	PDE11A	NM_001077196	unknown
chr4:42505466-42526864	+	ATP8A1	NM_001105529	hsa_circ_0069612
chr1:232596632-232607274	+	SIPA1L2	NM_020808	unknown
chr3:3197902-3215945	+	CRBN	NM_001173482	hsa_circ_0003400
chr5:145197456-145205763	+	PRELID2	NM_182960	hsa_circ_0006528
chr17:63685246-63746842	-	CEP112	NM_001037325	unknown
chr12:97886238-97954476	+	RMST	NR_024037	unknown
chr9:114860749-114875148	+	MIR3134	NR_036085	hsa_circ_0088046
chr1:62321701-62341038	+	PATJ	NM_176877	unknown
chr4:183522076-183575046	-	TENM3	NM_001080477	unknown
chr4:103225473-103236987	-	SLC39A8	NM_001135147	hsa_circ_0002782
chr3:89259009-89259670	-	EPHA3	NM_182644	hsa_circ_0066596
chr9:73399020-73479427	+	MIR204	NR_029621	unknown
chr21:27326903-27354790	+	APP	NM_001136016	unknown
chr3:196831773-196846401	-	DLG1	NM_001204387	hsa_circ_0008500
chr4:42487512-42526864	+	ATP8A1	NM_001105529	hsa_circ_0069608
chr1:65131739-65141666	-	CACHD1	NM_020925	hsa_circ_0007009
chr4:183245098-183268082	+	TENM3	NM_001080477	hsa_circ_0071475

chr1:62321701-62393501	-	PATJ	NM_176877	unknown
chr3:157839891-157921034	+	RSRC1	NM_001271834	hsa_circ_0067808
chr1:232649602-232651354	+	SIPA1L2	NM_020808	unknown
chr3:196817782-196846401	-	DLG1	NM_001204387	hsa_circ_0008583
chr4:42487512-42546003	+	ATP8A1	NM_001105529	unknown
chr2:171884848-171902872	+	TLK1	NM_001136555	hsa_circ_0004442
chr8:40532222-40554920	-	ZMAT4	NM_001135731	unknown
chr4:42487512-42509171	+	ATP8A1	NM_001105529	hsa_circ_0069607
chr1:65068488-65107652	-	CACHD1	NM_020925	unknown
chr17:63545637-63554854	+	AXIN2	NM_004655	hsa_circ_0045350
chr20:8720990-8746005	-	PLCB1	NM_015192	unknown
chr18:346294-357522	+	COLEC12	NM_130386	unknown
chr3:89444986-89468540	-	EPHA3	NM_182644	unknown
chr14:63447589-63483672	+	KCNH5	NM_139318	hsa_circ_0032148
chr9:74309424-74313120	+	CEMIP2	NM_001135820	hsa_circ_0003861
chr1:62321701-62367131	-	PATJ	NM_176877	unknown
chr3:77595488-77617585	-	ROBO2	NM_002942	unknown
chr2:171709223-171710532	+	GAD1	NM_000817	hsa_circ_0057012
chr9:115013208-115024879	-	MIR3134	NR_036085	hsa_circ_0088073
chr1:62455839-62516730	+	PATJ	NM_176877	unknown
chr4:107092251-107157965	+	TBCK	NM_001163435	hsa_circ_0070585
chr12:97856929-97954825	-	MIR1251	NR_031653	unknown
chrX:10534927-10535643	+	MID1	NM_001193277	hsa_circ_0007933
chr2:188243666-188252483	+	CALCRL	NM_001271751	unknown
chr11:128932174-128936763	+	ARHGAP32	NM_001142685	hsa_circ_0024840
chr1:62374043-62380336	-	PATJ	NM_176877	unknown
chr16:30740286-30745329	-	SRCAP	NM_006662	hsa_circ_0004236
chr2:206023445-206058044	-	PARD3B	NM_057177	hsa_circ_0008172
chr8:40532222-40532450	+	ZMAT4	NM_001135731	unknown
chr9:73376516-73426160	+	MIR204	NR_029621	unknown
chr1:62455839-62503721	-	PATJ	NM_176877	unknown
chr6:45459677-45480144	-	RUNX2	NM_001015051	hsa_circ_0076691
chr4:108984778-109000770	+	LEF1	NM_001166119	unknown
chr1:232551239-232568217	+	SIPA1L2	NM_020808	hsa_circ_0016909

Table 12: Pan-cohort circular RNA biomarker species for circular RNA-based SHH medulloblastoma group. Circular RNA information according to DCC.

coordinates	**strand**	**gene**	**refseqid**	**circ_id_circbase**
chr7:16298014-16317851	+	ISPD	NM_001101417	hsa_circ_0079480
chr4:119026172-119064839	-	NDST3	NM_004784	unknown
chr2:110321942-110350696	-	SEPT10	NM_144710	hsa_circ_0002076
chr8:18656804-18662408	+	PSD3	NM_206909	hsa_circ_0004458
chr5:64747301-64769779	-	ADAMTS6	NM_197941	hsa_circ_0072688

chr4:119026172-119036115	-	NDST3	NM_004784	unknown
chr10:86177526-86237420	-	CCSER2	NM_018999	hsa_circ_0018996

Table 13: Significant circular RNA biomarkers in Group 3 medulloblastoma (circular RNA medulloblastoma groups) in discovery and validation. Circular RNA information according to DCC.

discovery				
coordinates	**strand**	**gene**	**refseqid**	**circ_id_circbase**
chr15:76152218-76165909	-	UBE2Q2	NM_173469	unknown
chr2:227729319-227779067	-	RHBDD1	NM_001167608	hsa_circ_0058495
chr3:138289159-138290198	-	CEP70	NM_024491	hsa_circ_0002468
chr5:72370568-72373320	-	FCHO2	NM_001146032	hsa_circ_0002490
chr7:80418621-80440017	-	SEMA3C	NM_006379	hsa_circ_0004365
validation				
coordinates	**strand**	**gene**	**refseqid**	**circ_id_circbase**
chr1:216495224-216500996	+	USH2A	NM_007123	unknown
chr1:225140371-225195246	-	DNAH14	NM_144989	hsa_circ_0016601
chr1:231930987-231954263	-	unkn	NR_002227	hsa_circ_0007848
chr12:5841685-5916534	+	ANO2	NM_001278596	unknown
chr12:5860001-5941769	+	ANO2	NM_001278596	unknown
chr12:5908672-5941769	+	ANO2	NM_001278596	unknown
chr12:5915217-5941769	+	ANO2	NM_001278596	unknown
chr12:5936934-5941769	-	ANO2	NM_001278596	unknown
chr13:96577933-96651561	-	UGGT2	NM_020121	unknown
chr15:33445248-33447246	-	FMN1	NM_001277314	unknown
chr2:207144263-207162097	+	ZDBF2	NM_020923	hsa_circ_0002141
chr2:40366540-40405633	+	SLC8A1-AS1	NR_038441	unknown
chr2:40655612-40657441	+	SLC8A1	NM_001112800	hsa_circ_0005232
chr2:40655612-40657444	+	SLC8A1	NM_001112800	hsa_circ_0000994
chr3:18419661-18462483	-	SATB1	NM_001195470	hsa_circ_0064555
chr3:33725850-33738425	+	CLASP2	NM_015097	hsa_circ_0001280
chr3:68929880-68934461	+	FAM19A4	NM_001005527	hsa_circ_0066495
chr4:162376155-162431576	+	FSTL5	NM_001128427	unknown
chr6:135621637-135644462	+	AHI1	NM_001134830	hsa_circ_0005214
chr6:65300115-65303202	-	EYS	NM_001142800	unknown
chr6:66200486-66205886	+	EYS	NM_001142800	unknown
chr7:81662112-81746489	+	CACNA2D1	NM_000722	unknown
chr7:81689743-81746489	+	CACNA2D1	NM_000722	hsa_circ_0003159
chr8:104922361-104973361	-	RIMS2	NM_001100117	unknown
chr8:105080739-105161076	-	RIMS2	NM_001100117	hsa_circ_0005114
chr8:105105698-105161076	+	RIMS2	NM_001100117	hsa_circ_0085302
chr8:32453345-32474403	-	NRG1	NM_013962	hsa_circ_0007279
chr9:88284399-88327481	+	AGTPBP1	NM_015239	hsa_circ_0087391

Table 14: Pan-cohort circular RNA biomarker species for circular RNA-based Group 4 medulloblastoma group. Circular RNA information according to DCC.

coordinates	strand	gene	refseqid	circ_id_circbase
chr2:72945231-72960247	+	EXOC6B	NM_015189	hsa_circ_0009043
chrX:147743428-147744289	-	AFF2	NM_001169122	hsa_circ_0001947
chr4:39739039-39776553	+	UBE2K	NM_001111112	hsa_circ_0002590
chr10:32832227-32873232	-	CCDC7	NM_001026383	hsa_circ_0000233
chr2:72958135-72960247	+	EXOC6B	NM_015189	hsa_circ_0001030
chr13:78293666-78327493	-	SLAIN1	NM_001242868	hsa_circ_0000497
chr5:16779653-16783578	+	MYO10	NM_012334	unknown
chr10:49609654-49618211	-	MAPK8	NM_001278548	hsa_circ_0002968
chr14:50616725-50616948	+	SOS2	NM_006939	hsa_circ_0007695
chr4:73950965-73958017	+	ANKRD17	NM_032217	hsa_circ_0001417
chr6:17507399-17514185	-	CAP2	NM_006366	hsa_circ_0002245
chr6:70447833-70500364	+	LMBRD1	NM_018368	unknown
chr21:17553910-17603435	-	MIR99AHG	NR_027790	unknown

7.3 Figure directory

Figure 1: Medulloblastoma groups and subtypes displayed with characteristics of each subtype. Taken from Cavalli et.al. 2017..12
Figure 2: FastQC report summarized by MultiQC. Each horizontal line represents one sample and the mean quality (Phred) score in the corresponding position. The green area indicates generally good quality data, yellow acceptable and red bad quality. At least the two fastq files indicated by the yellow horizontal lines should to be discarded...................18
Figure 3: Per base sequence content plot from the same MultiQC report. Here the first few bases depict a clear shift into one of the four colors, indicating bad quality or adapter contamination. This adapter content will be removed before the analysis.........................19
Figure 4: Overview of circular RNA synthesis pathways. a) Linear DNA with introns and several exons. b) Back-splicing by complementary sequences flanking the exons. c) RNA binding protein (RBP) aided back-splicing. d) Exonic circular RNA with 2 exons. e) EIcircRNA with 2 exons and 1 intronic region. f) DNA with 1 exonic and 1 intronic region. g) Single-exon circular RNA synthesis by back-splicing and intron lariat formation. h) Single exonic circular RNA. i) Intronic lariat...21
Figure 5: Examples of circular RNA functions. a) MicroRNA sponging. b) RNA binding protein RBP sponging. c) Translation of a circRNA into a protein. d) Complex forming including a protein and a cofactor...22
Figure 6: The workflow of a typical CircleSeq protocol. The sample is split into two parts after RNA isolation and ribosomal RNA depletion. One of the two preparations is then treated with RNAseR while the other preparation is not. After sequencing both samples RNAseR- stable (circular RNAs that are enriched in the RNAseR treated sample) circular RNAs can be identified to confirm the presence of circRNAs in the untreated preparation. 24
Figure 7: Output of a circtools primer run. For each circular RNA, multiple potential primers are designed, giving the user a choice. Forward and reverse columns show the primer sequence, basic local alignment search tool (BLAST) reveals the uniqueness in the

genome. The GC% forward and reverse depict the Guanine and Cytosine (GC) content of the primers shown..25

Figure 8: The general data flow of *circs* including the main steps pipeline run, summary and annotation, vote and normalization. First all three pipelines were run on the same input data and filtered according to the minimal reads threshold (two back splice junction reads). Next these results were subsequently summarized and annotated with information from circbase, circbank, RefSeq and BioMart. In a third step, the output circular RNA candidates of all three pipelines were overlapped and only putative circular RNAs that were found in all three filtered output files were accepted. In a fourth step the DCC quantifications of accepted (voted) coordinates were normalized to junction reads per million total RNA-Seq reads...38

Figure 9: Correlation of raw circular RNA approved quantifications from the three datasets DCC, find_circ and CIRCexplorer1 correlated to each other. Linear trend line added. a) find_circ *versus* DCC. b) find_circ *versus* CIRCeplorer1. c)CIRCexplorer1 *versus* DCC. Each axis are raw circRNA junction quantifications...41

Figure 10: Circs_snake DAG. General rules are denominated by a number, workflow specific rules by dcc ,fc and cx for DCC, find_circ, CIRCexplorer respectively..................45

Figure 11: Amount of unique circular RNAs identified from each pipeline in the discovery cohort. DCC detected the most circular RNAs, find_circ an intermediate amount, while CIRCexplorer1 detected the least..49

Figure 12: The vote result of the discovery cohort. The vote including all three algorithms agreed on 8379 circular RNA species. The DCC quantifications of these circular RNA candidates were normalized and then used for all subsequent analysis...........................50

Figure 13: Heatmap of top 500 differentially expressed circular RNAs across all 38 samples in the discovery cohort. Top colors and dendrogram according to circs' clustering (average Pearson's dissimilarity). Blue= WNT medulloblastoma, red = SHH medulloblastoma, yellow = Group 3 medulloblastoma, green = Group 4 medulloblastoma. 51

Figure 14: Comparison of group classification from Forget et al. 2018 and top 500 differentially expressed circular RNAs (Pearson's dissimilarity) in the discovery cohort. Similarity network fusion (SNF) is a method using multiple input data layers, resulting in one consensus classification. When a sample lacked any omics data layer, SNF could not be performed. Blue = WNT medulloblastoma, red = SHH medulloblastoma, yellow = Group 3 medulloblastoma, green = Group 4 medulloblastoma, White boxes = no data available. 52

Figure 15: Circular RNA signal across medulloblastoma groups according to two differing MB group classifications. a) Circular RNA signal according in SNF MB groups. b) Circular RNA signal in MB groups according to circular RNA data. ***=Tukey HSD p-value adjusted <0.001. Normalized circRNA signal to back-splice reads per million.................................53

Figure 16: Heatmap of top 500 differentially expressed circular RNAs across all 35 samples in the validation cohort. Top colors and dendrogram according to circular RNA group classification (average Pearson's dissimilarity). Group colors according to discovery cohort legend (blue = WNT MB, red = SHH MB, yellow = Group 3 MB, green = Group 4 MB)..55

Figure 17: Circular RNA signal across medulloblastoma groups in validation cohort. Here the difference between Group 4 and SHH MB was significant.*= Tukey HSD p-value adjusted <0.05..56

Figure 18: Circular RNA signal across three datasets. ***=p-value adjusted Tukey HSD <0.001. Healthy = ENCODE healthy fetal brain tissue samples..57

Figure 19: Corrplot of discovery and validation circular RNA data. Pie color and filling represents Pearson correlation. Higher correlation is represented by a darker color and a more filled circle. Sample order according to circular RNA MB groups and cohort.............58

Figure 20: Group-specific biomarker identification in the discovery and validation medulloblastoma (MB) cohort. Significant circRNA expression differences according to MB group as determined by ANOVA followed by a Tukey HSD as a post-hoc test. a,b,c,d depict group-specific biomarker candidates in discovery and validation WNT, SHH, Group 3 and Group 4, respectively..59

Figure 21: Mean in-group expression of circular RNA biomarkers in discovery cohort versus validation cohort. Logarithmic scale of normalized circRNA junction reads per million. Names above or next to the data points show the parental gene of each circRNA candidate (if the circRNA consits of exons from geneA, the label is geneA)......................61

Figure 22: Boxplot of normalized circ*RMST* expression in the discovery cohort. Samples are ordered according to circular RNA medulloblastoma groups, genomic coordinates based on hg19. ***=Tukey HSD p-value adjusted<0.001..62

Figure 23: Megasampler of RMST RNA expression in several cancer data sets. Taken from https://hgserver1.amc.nl/cgi-bin/r2/main.cgi. CCLE=Cancer Cell Line Enzyclopedia, normalized expression values. A Megasampler is a boxplot comparing expression of one feature between multiple datasets...63

Figure 24: Boxplot of normalized circ*RMST* expression in the validation cohort. Samples are ordered according to circular RNA medulloblastoma groups, genomic coordinates based on hg19. ***=Tukey HSD p-value adjusted<0.001..63

Figure 25: Circ*ISPD* normalized expression in circular RNA defined medulloblastoma groups in discovery (a) and validation (b) cohort with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001...65

Figure 26: Circ*EXOC6B* normalized expression in circular RNA defined medulloblastoma groups in discovery (a) and validation (b) cohorts with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001...66

Figure 27: Circ*PATJ*, circ*EYA1*, circ*EYS* and circ*RPH3A* normalized expression in circular RNA defined MB groups in merged discovery and validation cohort combined with 12 fetal brain tissue controls with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001...68

Figure 28: Heatmap of the top 500 differentially expressed circular RNAs in the discovery (a) and validation (b) cohorts. Clustering according to Pearson's dissimilarity, color according to DNA methylation subtype information (see legend on the right side)............70

Figure 29: SHH CHLD AD subtype circular RNA biomarkers. Based on DNA methylation subtypes and combined cohort with genomic coordinates (hg19). Tukey HSD p-values adjusted: *<0.05,**<0.01,***<0.001..71

Figure 30: Circbase-based circular RNA validation of discovery cohort (a) and validation cohort (b).....73
Figure 31: MiOncoCircDB coordinates overlaps with discovery (a) and validation dataset (b). LiftOver was used to translate the hg19 coordinates of both datasets into hg38 coordinates, failed liftover coordinates were excluded from this analysis.....73
Figure 32: Overlap of discovery (a) and validation (b) cohort with all curated data bases. 74
Figure 33: Circ*RMST* in validation cohort and MiOnciCircDB data divided into four categories: Healthy_CNS, Healthy_Non_CNS, Cancer_CNS and Cancer_NON_CNS. All MiOncoCircDB samples with detected circ*RMST* were included in this plot.....75
Figure 34: MiOncoCircDB megasamplers for circ*EXOC6B* (a) and circ*ISPD* (b) in validation and MioOncoCircDB. N numbers refer to total samples in the validation cohort and circular RNA positive samples for the respective circular RNA in each plot.....76
Figure 35: Normalized circular RNA signal. a) for each sample, ordered by cell line and condition. b) across 4 conditions in 3 medulloblastoma cell lines combined. RNAseR= sample treated with RNAseR. ** = Tukey HSD p-value adjusted <0.01. MYC OE= MYC Overexpression.....77
Figure 36: Heatmap of top 500 differentially expressed circular RNAs across the *MYC*/RNAseR dataset. Normalized circRNA data. Clustering based on Pearson's dissimilarity, high circRNA expression = red, low circRNA expression = green.....79
Figure 37: Circbase - based overlap of circular RNA candidates included in the whole *MYC*/RNAseR dataset.....79
Figure 38: *MYC*/RNAseR samples and MiOncoCIrcDB overlap. All in the dataset included circular RNAs were used for this plot.....80
Figure 39: *MYC*/RNAseR samples and curated databases overlap.....80
Figure 40: RNAseR stable circRNAs found in the dataset relative to all non-RNAseR circRNA species across all three used cell lines.....81
Figure 41: RNAseR and non-RNAseR circular RNAs overlap in the *MYC*/RNAseR dataset. 82
Figure 42: *MYC*/RNAseR circRNA overlaps with discovery and validation cohort. Overlaps

of complete data (a) and Overlap of discovery (b) and validation (c) cohort with RNAseR-treated samples respectively....83

Figure 43: DAOY-specific RNAseR validation rate. R² based on a linear model through all data points....84

Figure 44: ONS76-specific RNAseR validation rate. R² based on a linear model through all data points....85

Figure 45: UW228-specific RNAseR validation rate. R² based on a linear model through all data points....86

Figure 46: Volcano plot of *MYC* overexpression specific circular RNA changes. red line signals statistical significance. the parental gene names of some circular RNAs are next to its datapoint....87

7.4 List of tables

Table 1: Consumables used for medulloblastoma cell line cultivation. 30

Table 2: software packages used in this thesis. 32

Table 3: Non-default Perl packages used during this project. 32

Table 4: Non-default R packages used in this thesis. 32

Table 5: Hardware utilized during this thesis. 33

Table 6: Comparison of DNA methylation subtypes in the discovery and validation cohorts. 71

Table 7: Circular RNA signal across all *MYC*/RNAseR samples. Circ signal decrease calculated in percent of baseline, *MYC* OE effect calculated based on x times baseline.....78

Table 8: *In silico* validation rate of the three MB circRNA datasets included in this thesis. 80

Table 9: Table of RNA-Seq preparation metrics for the MYC/RNAseR dataset used in this thesis 109

Table 10: RNA-Seq MYC quantifications across MYC/RNAseR dataset. 109

Table 11: Pan-cohort circRNA biomarker species for circ medulloblastoma RNA-based WNT medulloblastoma group. Circular RNA information according to DCC. 109

Table 12: Pan-cohort circular RNA biomarker species for circular RNA-based SHH medulloblastoma group. Circular RNA information according to DCC. 111

Table 13: Significant circular RNA biomarkers in Group 3 medulloblastoma (circular RNA medulloblastoma groups) in discovery and validation. Circular RNA information according to DCC 112

Table 14: Pan-cohort circular RNA biomarker species for circular RNA-based Group 4 medulloblastoma group. Circular RNA information according to DCC. 113

Curriculum Vitae

Personal Data

Name: Daniel Rickert

Address: Eulerstraße 31,40447 Düsseldorf

Email: danielrickert@protonmail.com

Date/Place of Birth: 31.03.1993/ Borken, NRW

Education

2018-present	PhD Bioinformatics at Pediatric Neurooncology UKD Düsseldorf “circRNA in Medulloblastoma“ Prof. Dr. Marc Remke, Prof. Dr. Guido Reifenberger, Prof. Dr. Gunnar Klau
2015 – 2018	Master Biology (Major: quantitative Biology and Bioinformatics) at Computational Cell Biology Institute HHU Düsseldorf: “Predicting enzyme and metabolite concentrations in the core metabolism of E. coli from rate laws and cost minimization“ Prof. Dr. Martin Lercher, Prof. Dr. Oliver Ebenhöh
2012 – 2015	Bachelor Biology at Computational Cell Biology Institute HHU Düsseldorf: “Does metagenomic data contain older relatives of orphan genes?“ Prof. Dr. Martin Lercher, Prof. Dr. Gerhard Steger
2009 – 2012	A-Levels at Hans-Böckler-Berufskolleg, Marl combined with a BTA - apprenticeship (biological technical assistant)

Internships

2012	Hygiene Laboratories at Wilhelms-Universität Münster
	Human genetics Laboratories at Wilhelms-Universität Münster
2011	Microbiology Laboratories at Iglo
	MRSA Laboratories Illinois State University, USA

Work Experience

2018	WHK at DKFZ / DKTK Essen Düsseldorf (AG Remke)
	automating circular RNA detection from RNA-Seq pipelines
2015 – 2018	WHK at Computational Cell Biology Institute HHU Düsseldorf
	Work on metagenomic data utilizing Perl, R, BLAST, hmmer

Publications and presentations

Publications included in this thesis

Rickert D, Bartl J, Picard D, et al. Circular RNA profiling distinguishes medulloblastoma groups and shows aberrant RMST overexpression in WNT medulloblastoma. Acta Neuropathol. 2021;141(6):975-978. doi:10.1007/s00401-021-02306-2

Poster and oral presentations

CircRNA in Medulloblastoma.
Daniel Rickert, Jasmin Bartl, Daniel Picard, Flavia Bernardi, Nan Qin, Marta Lovino, Stéphanie Puget, Frauke-Dorothee Meyer, Ute Fischer, Arndt Borkhardt, Guido Reifenberger, Olivier Ayrault, Marc Remke. Virtual poster presentation, 4th **Translational Oncology Symposium Virtual, 18.02.2021**
CircRNA in Medulloblastoma.
Daniel Rickert, Jasmin Bartl, Daniel Picard, Flavia Bernardi, Nan Qin, Marta Lovino, Stéphanie Puget, Frauke-Dorothee Meyer, Ute Fischer, Arndt Borkhardt, Guido Reifenberger, Olivier Ayrault, Marc Remke. Presentation, 3rd **Translational Oncology Symposium Goch, 06.02.2020**
CircRNA in Medulloblastoma.
Daniel Rickert, Jasmin Bartl, Daniel Picard, Flavia Bernardi, Nan Qin, Marta Lovino, Stéphanie Puget, Frauke-Dorothee Meyer, Ute Fischer, Arndt Borkhardt, Guido Reifenberger, Olivier Ayrault, Marc Remke. Poster, 2nd **Translational Oncology Symposium Essen, 07.02.2019**

Acknowledgments

First and foremost I would like my Supervisors Prof. Dr. Marc Remke, Prof. Dr. Guido Reifenberger and Prof. Dr. Gunnar Klau for envisioning, supervising and and helping out with the project whenever needed.

I would like to thank the Düsseldorf School of Oncology for financial Support.

To a greater extend, a big thank you goes out to the AG Remke, former and current members for support, new ideas and all the shared group activities, inside and outside the UKD.

Daniel Picard earned special acknowledgment: emotional support, political advisor and always having interesting stuff for me to do, even if it is again helping you to get some fringe software running or keep on running. You helped me enormously throughout the years- Thank you.

Coauthors of the here included publication and the whole project did earn their place here aswell, thanks for taking part in each round of improving the science presented here. I would like to thank the KMT lab members aswell for the good times spend together and advice given. Thanks to Prof. Dr. Ulf Kahlert, the person who wanted to employ me – which is why I ended up in the AG Remke and even after was willing to help whenever he could. The Bioinformaticans Dr. Deya Alzoubi, Dr. Layal Yasin, Phillip Spohr helped me fixing the many bugs I encountered, thank you!

Thanks also to computational support of the Zentrum für Informations- und Medientechnologie, especially the HPC team (High Performance Computing)

at the Heinrich Heine University Düsseldorf.

Alex, thank you for your encouragement, help, support and for just being there for me throughout the years.

Finally, the biggest thank you goes to my family, the most important part of getting me where I am now.

Affirmation

Hereby, I declare on oath that I composed this dissertation independently by myself. I used only the references and resources indicated in this thesis. With the exception of such quotations, the work presented in this thesis is my own. I have accredited all the sources of help. This PhD thesis was never submitted or presented in a similar form to any other institution or examination board. I have not undertaken a doctoral examination without success so far.

____

Daniel Rickert