Year : 2017

Recently active transposable elements provide insights into the evolution of mammalian circular RNAs

Gruhl Franziska

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Centre intégratif de génomique

Recently active transposable elements provide insights into the evolution of mammalian circular RNAs

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Franziska GRUHL

Master de l'Université de Heidelberg, Heidelberg, Allemagne

Jury

Prof. Nicolas Perrin, Président
Prof. Ioannis Xenarios, Directeur de thèse
Prof. Henrik Kaessmann, Co-directeur de thèse
Prof. assistant boursier FNS Vincent Dion, expert
Prof. Mark Robinson, expert

Lausanne 2017

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président· e	Monsieur	Prof.	Jérôme	**Goudet**
Directeur· rice de thèse	Monsieur	Prof.	Ioannis	**Xenarios**
Co-directeur· rice	Monsieur	Prof.	Henrik	**Kaessmann**
Experts· es	Monsieur	Prof.	Vincent	**Dion**
	Monsieur	Prof.	Mark	**Robinson**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Franziska Gruhl

Master of Science at University of Heidelberg, Allemagne

intitulée

Recently active transposable elements provide insights into the evolution of mammalian circular RNAs

Lausanne, le 17 juillet 2017

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Jérôme Goudet

Contents

List of Figures

List of Tables

Abstract

The transcriptional landscape of the mammalian genome consists of a variety of different RNAs, such as protein-coding RNAs (mRNAs), long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) among others. Alternative splicing further diversifies the observed landscape. However, a high number of alternatively spliced transcripts is produced by errors during transcription and splicing and is referred to as transcriptional noise without functional significance.

Circular RNAs (circRNAs) constitute a class of RNAs that was only recently discovered. Little is known about their properties and functional relevance in respect to other RNAs of the transcriptome. Moreover, it is unknown if - as claimed in several studies - they are beneficial to the organism and thus have been selectively retained in the genomes of a variety of species.

By analyzing circRNAs across a set of five mammalian species representing different mammalian lineages, I provide first evidence that circRNAs are predominantly a transcriptional by-product caused by the integration of species-specific and recently active transposable elements (TEs).

CircRNAs are expressed at low levels. Their biogenesis is influenced by TEs in the flanking introns of the circRNA causing the formation of a hairpin structure in the pre-mRNA that allows backsplicing. The integration and fixation of TEs in coding genes is biased to genes that are GC low and have a complex structure (many exons, long introns) leading to a subset of genes predisposed to produce circRNAs. The independent targeting of structurally similar genes by TEs has led to the independent emergence of circRNAs in orthologous genes of multiple species. The tight link between circRNA expression and recently active, species-specific TEs suggests that many circRNAs are transcriptional noise.

Résumé

Le paysage transcriptionnel des génomes mammifères consiste en une variété d'ARNs différents, tels que, entre-autres, les ARN codant pour des protéines (ARNm) et les micro-ARNs (miARN). L'épissage alternatif diversifie encore plus le paysage observé. Pourtant, un grand nombre de transcrits venant de l'épissage alternatif est dû à des erreurs pendant la transcription et l'épissage, et sont qualifiés de bruit transcriptionnel sans signification fonctionnelle.

Les ARNs circulaires (ARNcircs) constituent une classe d'ARNs qui n'a été découverte que récemment. Leurs propriétés, de même que leur pertinence fonctionnelle, sont peu connues en comparaison des autres ARNs du transcriptôme. De plus, on ignore encore si – comme le prétendent certaines études – ils sont bénéfiques pour l'organisme, et par conséquent ont été sélectivement conservés dans le génome d'un certain nombre d'espèces.

Grâce à l'analyse des ARNcircs au travers d'un ensemble de cinq espèces de mammifères représentant plusieurs lignées, je fourni les premières indications que les ARNcircs sont de manière prédominante des sous-produits provenant de l'intégration d'éléments transposables propres à l'espèce et récemment actifs.

Les ARNcircs sont faiblement exprimés. Leur biogenèse est influencée par les éléments transposables dans les introns adjacents de l'ARNcirc, causant la formation d'une structure en épingle-à-cheveux qui permet le rétro-épissage. L'intégration et la fixation des éléments transposables dans les gènes codants sont biaisés en faveur des gènes qui ont un faible contenu en GC, et qui ont une structure complexe (beaucoup d'exons et de longs introns), qui constituent donc un sous-ensemble de gènes prédisposés à produire des ARNcircs. Le ciblage indépendant de gènes similaires par les éléments transposables a amené l'émergence indépendante d' ARNcircs dans les gènes orthologues chez plusieurs espèces. Le lien étroit entre l'expression d'ARNcircs et les éléments transposables propres à l'espèce récemment actifs suggère que de nombreux ARNs circulaires sont des sous-produits de bruit transcriptionnel.

Acknowledgements

I would like to thank Henrik Kaessmann for the opportunity to work as a PhD student in his lab, for allowing me to explorer and to shape my PhD project in different directions and for all the scientific discussions. It has been an interesting experience and I am grateful that I could make this experience and learn from it.

Furthermore, I would like to thank Ioannis Xenarios for accepting me as a PhD student in Vital-IT. I am grateful for all your support and for getting to know the working atmosphere and people in Vital-IT.

Thank you also to the remaining members of my thesis committee - Nicolas Perrin, Mark Robinson and Vincent Dion - for accepting being part of the committee and for the scientific discussion.

I would like to thank David Gatfield for this great collaboration. Thank you, for giving me the opportunity to work in your lab, for the constant interest in the project and all the interesting questions and ideas.

Great thanks to all my colleagues - from the old lab in Lausanne, the new lab in Heidelberg and from Vital-IT. It has been a pleasure to work together with so many different people from different backgrounds and to constantly get new input and ideas on my project. I am specifically grateful to Konstantin Popadin, Michael Saina, Mihai Petrovici and Iris Finci for reading, correcting and discussing my thesis manuscripts with me. Thank you to Tania Studer and Maxime Jan for taking care of the French translations.

I would also like to say thank you to the SIB for awarding me with a PhD fellowship in the first place (provided by the Leenaards foundation), but also for offering all the different training courses in Bioinformatics, which have been of a great help for me.

Last, I would like to thank all my friends and especially my family, who have supported me throughout the last years.

1 Introduction

The transcriptional landscape of the mammalian genome consists of a variety of different RNAs, such as protein-coding RNAs (mRNAs), long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) among others. Together, they form and orchestrate various networks that are not only maintaining every-day cell and tissue homeostasis, but are also responsible for the developmental processes that are transforming a single-cell zygote into a complex living organism. Each transcript is characterized by its own spatio-temporal expression profile. Some of these expression profiles are well conserved between species, while others reflect species-specific adaptations. The rise of next-generation sequencing approaches led to a dramatically increased number of annotated RNAs and expression profiles, which has challenged our understanding of species complexity. However, the expression of a high number of these transcripts might only have little or no impact on the organism. It is considered transcriptional noise, which will only experience weak selective pressure if at all. [1]. The expression of deleterious transcript variants in contrast, is subject to purifying selection - a process, in which deleterious variation is removed.

Circular RNAs (circRNAs) constitute a class of RNAs that was only recently described in large-scale in mammalian genomes. Although found in the majority of analyzed genomes, little is known about their contribution to the observed species complexity. To understand whether the expression of circRNAs is an important and conserved feature of mammalian genomes, one needs to analyze them in the context of the whole genome and its properties. In the introductory chapters, I will therefore provide an overview of different processes that shape the mammalian genome, before discussing the characteristics of circRNAs. I will start by introducing the concept of isochores to show how genomes are globally structured. I will then continue by explaining the processes of transcription and splicing, which result in the observed complexity of the mammalian RNA landscape. The transcription of DNA and splicing of mRNA are stochastic processes and I will describe their robustness in light of transposable elements. Finally, I will summarize the state of the art of the circRNA research field and discuss some of the current working hypotheses.

1.1 Mammalian genomes and transcriptomes

1.1.1 Isochores

Avian and mammalian genomes can be divided into adjacent DNA stretches ($>$ 300 nt) that differ from each other in their local base composition (proportion of A, T, G and C), but are fairly homogeneous on their own. The discovery gave rise to the term isochore and the grouping of

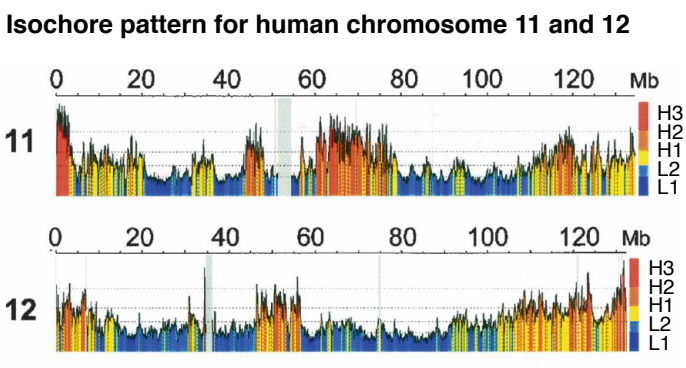


Figure 1: Isochore landscape for human chromosome 11 and 12 | GC content is plotted along the whole length of chromosome 11 and 12. Each vertical bar represents an isochore. The GC content in each isochore is color-coded as follows: L1 < 37%, L2 = 37-41%, H1 = 41-46%, H2 = 46-53%, H3 > 53%. *Figure adapted from Figure 1 in Costantini et al. [8].*

genomic regions into distinct isochores based on their GC content: L1 (< 37%), L2 (37-42%), H1 (42-47%), H2 (47-52%) and H3 (> 52%) (**Figure 1**, reviewed in [2]). Isochores are associated with distinct features of gene structure, function and expression. Amit *et al.* showed that exons can be separated into two groups based on their GC content and length of the flanking introns. GC-poor exons are flanked by large introns, while GC-rich exons are flanked by short introns. Furthermore, the GC difference between GC-poor exons and their introns is large (the GC difference between exons and introns is also known as GC amplitude), whereas GC-rich exons are accompanied by introns of similar GC content. In addition, Amit *et al.* suggested that the strong GC amplitude observed in GC-poor exons helps the splicing machinery to recognize these exons and to position itself correctly onto them ("exon definition mode"). In contrast, short, GC-rich exons are spliced based on "intron definition" (see **Chapter 1.1.2** for further details). In both categories, mutations that interfere with the local GC pattern lead to mis-spliced transcripts [3]. Several studies provided evidence that the local GC content is important for nucleosome positioning, methylation patterns and genome structure [4–6]. Interestingly, tissue-specific genes tend to be in GC-poor isochores, while broadly expressed genes are enriched in isochores with elevated GC levels [7]. Furthermore, different isochores are associated with different gene ontology (GO) terms. While L1 and L2 are overrepresented with functional gene categories related to sensory perception and processing, genes of intermediate GC levels (L2-H2) tend to have housekeeping functions (molecular binding and organization). GC-rich genes (H2/H3) are enriched in developmental processes and DNA-binding GO categories [7].

The genomic isochore patterns are largely conserved between species [3]. The ancestral genome, as proposed by Galtier *et al.*, was GC-poor, but already possessed regions with distinct GC content. Throughout time, GC differences became more distinct in all mammalian groups, except for rodents. The latter present an evolved state in which the GC variability between isochores has decreased [9].

It is still a matter of debate how and why isochores evolved. Current theories can be grouped into two major causes: Neutral evolution and selection. Supporters of the "neutral evolution cause" argue that isochores are merely a by-product of the mutation bias during DNA repair and GC-biased gene conversion. In contrast, proponents of the "selection cause" argue that the high GC content of mammalian genomes evolved as a consequence of the changes in homothermic body temperatures. Selective pressure acted on high GC content, because GC-rich DNA is thermodynamically stable and bendable. DNA stability and flexibility influence chromatin conformation and allow regulatory flexibility in warm-blooded organisms. For each theory, different pro and contra arguments exist, and it is difficult to find a common explanation (reviewed in [10]).

Besides the remaining controversies about the origin of isochores, it is evident that the local GC content needs to be taken into account when studying gene structure, expression and splicing.

1.1.2 DNA transcription and splicing

In a classical and simplified view, a coding gene consists of a promoter, exons and introns. Exons hold the functional information encoded by the gene, whereas introns typically contain regulatory signals. During transcription, the DNA is transcribed by an RNA polymerase into a pre-mRNA, which contains both exons and introns. In the next step, the different exon parts are joined to each other creating mRNA - a reaction which is known as splicing and is catalyzed by the spliceosome. Finally, the mRNA is translated into a protein.

DNA transcription can be divided into three major steps: Initiation, elongation and termination. During the initiation phase, the RNA polymerase binds together with additional transcription factors to the promoter of a gene. The DNA is unwinded and the elongation phase starts. During this process, the polymerase traverses along the DNA strand and forms the pre-mRNA by complementary base pairing with the DNA. Once the polymerase acquires a termination-sequence at the end of the gene, transcription stops. The pre-mRNA is released and is now ready for splicing.

The spliceosome is a complex of small nuclear rinbonucleoproteins (snRNPs). During the splicing reaction, the 5'-end of an intron (donor) is joined to the 3'-end of the intron (acceptor) and in a series of subsequent reactions, the intron is removed and the two exons are connected. The spliceosome recognizes exons and introns based on two mechanisms. In the first mode, the spliceosome uses introns as splicing unit ("intron-defined" mode). It positions the basal splicing machinery across introns to splice them out. In the second mode, the spliceosome identifies exons as splicing unit ("exon-defined" mode). The basal splicing machinery is now positioned on exons and they are spliced

together. Intron-defined splicing represents the ancestral state [11]. However, the capability of the splicing machinery to detect introns correlates negatively with the intron length. As a consequence of increased intron length throughout mammalian evolution, the exon-defined splicing mechanism evolved [3]. Alternative splicing describes a process in which exons of a gene are included, excluded or shuffled, thus leading to a situation in which a single gene can produce multiple, distinct transcripts. These transcripts differ in their base pair composition and have the potential to fulfill different tasks.

DNA transcription and splicing are biochemical reactions. Like any other biochemical reaction, they are subject to stochasticity depending on the local concentration of proteins and temperature. However globally, gene expression is a robust process that creates similar expression profiles across different species. To understand how a stochastic process can be robust at the same time, one needs to understand the different sources causing and controlling stochasticity.

1.1.3 Stochasticity during DNA transcription

RNA and protein abundance fluctuates within cells of the same tissue, because the biological mechanisms that are producing RNAs and proteins in a cell are stochastic in nature. The observed differences are often referred to as noise and are part of each biological system, independent of its scale. Stochasticity is an important factor to create heterogeneous cell populations allowing them to react rapidly to changes in the environment. Each gene has its own characteristic source and amplitude of noise, which is classified by its origin: Intrinsic noise can arise from any intracellular, biochemical reaction (e.g. DNA methylation, transcription, splicing or translation), while extrinsic noise is created by environmental changes (e.g. developmental transitions, stress) [12, 13].

The most prevalent source of intrinsic noise is mRNA transcription. In eukaryotes, transcription occurs in bursts that can differ in frequency and amplitude from each other [14, 15]. Genes in genomic regions with low transcriptional activity increase burst frequency to modulate their expression levels. The overall expression in a cell population may thus be stable, but different cells can be in different phases (before, in or after a transcriptional burst), which leads to high cell-to-cell variation. Genes in active regions are more likely to adapt the burst amplitude in order to change their activity [14]. They look less noisy between cells of the same cell population, because of a constant expression base line. Importantly, fluctuations decrease with the expression strength of a gene [16], meaning that higher noise levels are usually found in genes with lower expression levels. High variability in transcriptional levels can be buffered by translation efficiency. Therefore, proteins and their relative frequencies fluctuate less between cells than mRNAs (**Figure 2**) [13].

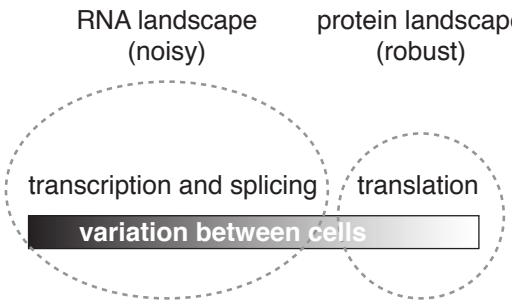


Figure 2: RNA and protein variability | RNAs and their absolute frequencies vary strongly between cells. Transcriptional differences can be buffered by translation rates and thus protein landscapes are less noisy than the preceding RNA landscapes between cells.

Martinez Arias *et al.* proposed that genes can be grouped into three distinct classes, each different in its properties of noise and "memory": Class I genes are either active or not. They respond rapidly to metabolic changes or stress conditions. As their task is merely to react or not, they do not require any memory of the cellular state. Expression can thus be noisy, and noise might even help to react differently at different points of time. Class II genes are necessary to keep systems in a stable, long-lasting state (e.g. differentiated tissues). Genes of this class have some kind of memory of the cell state. Their expression levels are stable and they do not allow high variation. Class III genes are intermediates, which keep a state only for a small period of time [17].

Fluctuations in gene expression levels between cells of the same tissue can be for stochastic reasons. In addition, transcripts with low expression levels harbor higher noise levels challenging the assessment to what extent their expression profiles are conserved across species: Expression shifts could either reflect adaptive changes and selection, or random shifts in the burst frequency and amplitude. Stochasticity in DNA transcription is thus an important factor to be taken into account when analyzing gene expression profiles and their conservation between species.

1.1.4 Stochasticity during mRNA splicing

Approximately 90% of human genes are subject to alternative splicing and the number of detected transcripts correlates strongly with the number of exons [1]. The average gene has four exons and thus needs to undergo three splicing reactions to form the final mRNA [18]. Based on the overlap of splice sites from the major isoform with others, Pickrell *et al.* suggested that in a coding gene an intron has a 0.7% probability to be mis-spliced. Per gene, this adds up to approximately 2% of mis-spliced transcripts. Following this logic, genes with more exons and higher expression should show a higher error rate in splicing. In contrast, Pickrell *et al.* found that genes with more exons and higher expression are subject to fewer splicing errors. They hypothesized that these genes are

under selection to keep fluctuations small, as an overproduction of transcripts could lead to a toxic accumulation of them [18].

It is important to understand that the splice site availability in a gene is not merely a function of its exon number. There are several other parameters that can determine how well or noisy splicing proceeds. For instance, the secondary structure of the mRNA can influence its availability for splicing [15, 19]. Furthermore, the presence of multiple, alternative splice sites can interfere with the recognition of the correct splice site. The more splice sites, the more likely an error can happen. Alternative splice sites can either evolve by mutation or be introduced by the integration of a transposable element (latter reviewed in [20]). Genes with higher mutation or transposon-integration rate are therefore more likely to produce mis-spliced transcripts. The integration of repeats in close proximity to an exon boundary can influence the local GC content. GC-rich SINE elements for example that integrate in close proximity to a splice site can lead to a local increase of GC, which decreases the GC amplitude at the exon-intron boundary. Especially in GC-low genes, this can interfere with the intron-defined mode of splicing and cause mis-splicing (see **Chapter 1.1.2**) [3].

Alternative transcripts may be grouped into different classes based on their presumptive purpose: 1) Transcripts that exhibit a function, 2) non-functional transcripts that merely by their presence interfere with the total number of functional transcripts, 3) non-functional, but constantly produced transcripts and 4) non-functional, stochastically produced transcripts [1]. All transcript types contribute to the observed RNA landscape, but they are subject to different evolutionary pressure. Melamud *et al.* proposed a model in which the the majority of transcripts are a product of noisy splicing. Some genes allow for noisy splicing and will posses many alternative isoforms, while other genes are under more constrain to keep the number of splicing errors low [1].

Stochasticity occurs at any level of biological systems as demonstrated by the examples of DNA transcription and mRNA splicing. However, the majority of noise is not harmful. Therefore, noise should not be seen as a source of error, but rather as *"[...] a landscape of opportunities in which novel biological activity can be explored at little cost."* [1]

1.1.5 Mammalian transcriptome diversity and conservation

The transcriptome of a given species reflects the expression levels of a variety of different coding and non-coding transcripts. Expression levels can be strong or weak, broad or tissue-specific and continuous or timed. Different organs distinguish themselves from each other by the number of transcribed

genes (coding and non-coding). The testis for instance, is characterized by a widespread expression of many genes facilitated by permissive (open) chromatin. Many of the expressed transcripts are poorly conserved and are likely non-functional. Liver in contrast, is characterized by a low, although strongly expressed number of genes [21].

Each gene has its own spatio-temporal expression pattern characteristic for the development and homeostasis of the species it is expressed in. Some of these patterns are conserved between species, others are not. The expression levels of orthologous protein-coding genes for instance, exhibit low variation between homologous tissues of different species. Interestingly, tissues are characterized by different changes in gene expression rates. Neuronal tissues evolve very slowly leading to a strong correlation of protein-coding gene expression levels between distant species. Liver, kidney and heart evolve at moderate levels, while testis is the most rapidly evolving tissue [22]. The differences in expression divergence rates might be explained by functional properties of the organ: The brain is under strong constrains to maintain the organism's integrity, whereas liver plays an important role in metabolic control and is more likely to reflect species-specific adaptations. Testis is subject to an intense and species-specific sex-related selective pressure, leading to its high rate of transcriptome evolution ([21], reviewed in [23]).

In contrast to the conserved expression profiles of coding genes, alternative splicing patterns evolve at much higher rates and are species-specific [23, 24]. As hypothesized by Melamud *et al.*, the majority of alternative isoforms originates as a by-product from noisy splicing, which is in line with the high number of species-specific alternative transcripts and their high divergence rates [1].

Transposable elements (TEs) are one source influencing alternative splicing patterns. They often possess species-specific amplification rates, which is reflected by the complex repeat landscapes in different organisms. In the next chapters, I will therefore discuss the properties of different transposable elements to illustrate how they can drive the formation of new isoforms.

1.2 Transposable elements

Mammalian genomes are rich in repetitive structures, which originate either from small, sequentially arranged DNA fragments, or from transposable elements spread throughout the genome. Depending on the organism, repetitive structures make up 40-50% of the genome, with the majority of repeats being derived from TEs (RepeatMasker, Feb 2017). Based on their mode of propagation, TEs are classified into retrotransposons (class I) and DNA transposons (class II). DNA transposons amplify via a "cut-and-paste" mechanism. They encode their own transposase, which can cut and re-integrate

the TE into the genome (**Figure 3A**). Retrotransposons propagate by reverse-transcription of an RNA intermediate, which is re-integrated into the genome. They often encode their own reverse transcriptase (RT) and integrase, but can also use the transcription and integration machinery of other TEs to do so. In addition, retrotransposons are divided into two subclasses: Long-terminal repeat (LTR) retrotransposons and non-LTR retrotransposons. LTR transposons are flanked by long, terminal repeats that bind to homologous regions in the genome and facilitate integration by recombination (**Figure 3B**). Non-LTR retrotransposonsons generate small DNA breaks that trigger reverse transcription of the RNA intermediate and subsequent integration into the genome (**Figure 3C**) [25].

1.2.1 Long interspersed elements

Long interspersed nuclear elements (LINEs) are non-LTR retrotransposons that resemble retroviruses. They constitute the largest class of transposable elements in human, but also in many other mammalian genomes (RepeatMasker, Feb 2017). Several subgroups such as L1, L2 or L3 exist. The L1 element is the most recent and only LINE element that is still active in the human genome. It consists of a 5'-UTR containing an RNA polymerase II promoter, two ORFs (open reading frame) that encode an RNA binding protein as well as an endonuclease (RNase H)/RT and a 3'-UTR harboring a poly(A)-tail. The full L1 element has a size of approximately 6 kilobase pairs (kb) [26]. In the human genome, 17.5% of detected transposons belong to the L1 family. However, the majority of L1 elements (99%) is degraded and has an average size of only 0.9 kb [27]. It is estimated that in human, one new integration per 20-200 births occurs [26]. L1 elements are not randomly distributed in the genome. They are abundant in AT-rich, gene-poor and weakly recombining regions. Younger L1s are closer to genes than older L1s. Additionally, they tend to occur in genes with lower expression levels [28–30]. The consensus sequence for the L1 endonuclease is TT|AAAA - an AT-rich motif, which might explain why L1s are more frequently found in AT-rich regions (discussed in **Chapter 1.2.3**). The human genome harbors intergenic L1 integration hotspots, but until now, no commonalities were found between them [30]. L1s can also mobilize various elements such as other TEs (SINEs), small non-coding RNAs or mRNAs and thus play an important role in shaping our genome (reviewed in [31]).

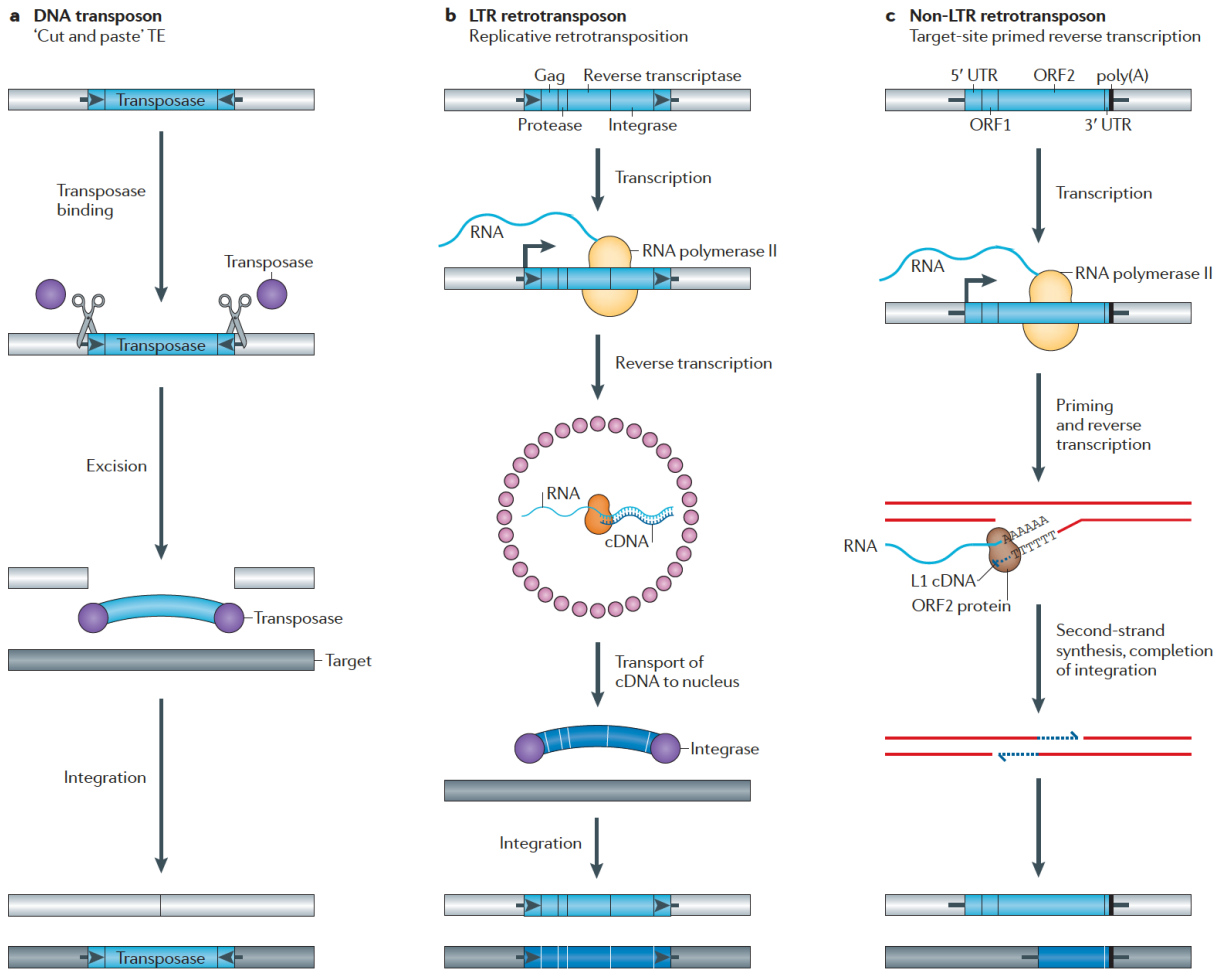


Figure 3: Mechanisms of transposon mobilization | A: DNA transposons propagate by a cut-and-paste mechanism, in which a transposon-encoded enzyme - the transposase - cuts the the transposon from the DNA and re-integrates it into the genome. **B:** LTR retrotransposons use RNA intermediates that are reverse-transcribed and re-integrated into the genome by the help of long-terminal repeats. **C:** Non-LTR retrotransposons generate small DNA breaks, which trigger reverse-transcription and integration of the RNA intermediate. *Figure adapted from Figure 1 in Levin and Moran 2011 [25].*

1.2.2 Small interspersed elements

Small interspersed elements (SINEs) are nonautonomous non-LTR retrotransposons. They are transcribed by RNA polymerase III (Pol III) and use the reverse-transcription and integration machinery of other TEs (mainly LINEs) to propagate throughout the genome. SINEs consist of a head, a body and a tail spending together over 0.1-0.6 kb. The head originates from small, Pol III-transcribed RNAs (tRNA, 7SL RNA, 5S rRNA). The SINE body shares homology with L1 elements, but it is unclear if it originated from L1 or independently. The sequence similarities allow SINEs to use

Table 1: Repetitive elements in different genomes | The frequency of repetitive elements in opossum, mouse, rat, rhesus macaque and human is summarized. Indicated is the fraction of each genome composed of repetitive elements, the fraction of each genome which is derived from a TE, frequent SINE families and recently active SINE families (number of detected TEs in brackets). For rhesus macaque, the percentage of the genome that is TE-derived was not found in literature. The frequency of the individual TEs was estimated with the RepeatMasker annotation as of Feb 2017.

Species	% Repetitive elements	% TE-derived	Frequent families	Recently active	Source
Opossum	56.6	52.0	RTESINE, SINE1_Mdo, SINE1a_Mdo	SINE1_Mdo (530,000)	Nilsson *et al.* [33] Nilson *et al.*, [34] Vassetzky *et al.*, [35]
Mouse	45.0	37.0	B1, B2, B4	B1 (570,000)	Kramerov *et al.*, [32] Vassetzky *et al.*, [35]
Rat	42.5	40.0	B1, B2, B4	ID (190,000)	Kim *et al.*, [36] Kim *et al.*, [37]
Rhesus	49.3	?	Alu	AluY (125,000)	Rhesus consortium, [38] Han *et al.*, [39]
Human	52.2	45.0	Alu	AluY (146,000)	Lander *et al.*, [40]

the L1 machinery for their own propagation. The SINE tail is build from small repeats, through which SINEs can concatenate and form dimers or trimers. Because SINEs rely on the L1 proteins for amplification, they show a similar integration bias as L1s into AT-rich regions [32] (discussed in **Chapter 1.2.3**).

SINE families have occurred at least 23 times independently in placental mammals. The evolution of a novel SINE family depends on several steps: First, an existing Pol III-transcribed RNA needs to be pseudogenized. Pseudogenization consists of the reverse-transcription of the RNA molecule into DNA followed by integration into the genome. Next, the pseudogenized gene needs to be transcribed at a developmental time, in which the genome is susceptible for TEs. Furthermore, the pre-SINE element needs to develop structures homologous to the RT of LINE elements to be recognized and efficiently transcribed and integrated. Last, changes in the secondary structure of the new SINE need to occur to avoid interference with the original pathway of the pseudogenized small RNA. As a consequence, LINE and SINE elements often co-evolve and show correlated activity and integration patterns [32]. This evolutionary trajectory often leads to species-specific SINE elements and closely related species can exhibit different SINE activity schemes. **Table 1** provides an overview of the frequency of repetitive elements in the organisms used in this study (opossum, mouse, rat, rhesus macaque, human). In the subsequent part, I will briefly summaries the species-specific evolution of SINE elements in each of them.

Opossum

With more than 530,000 copies SINE1_Mdo is the most frequent SINE element in opossum. It is still active [33]. In contrast to other elements, SINE1_Mdo has a complex structure with a head, core and tail domain. It uses L1 elements to propagate, but the exact integration and fixation processes have not been studied in detail (**Figure 4A**) [33–35].

Mouse

B1 is the most common SINE family in mouse (570,000 copies). It originated from a 7SL RNA and developed into a quasidimer that consists of the 7SL RNA, followed by a region of internal duplications of 20-30 nucleotides (nt). B1 TEs use the amplification machinery of L1. They have been very active in the mouse lineage (**Figure 4B**) [32, 35].

Rat

The ID family has undergone a recent and strong amplification round in the rat genome ($>$ 190,000 copies in rat versus 64,000 in mouse, RepeatMasker as of Feb 2017). The ID family originated from a tRNA. One of the earliest ID-elements was the BC1 RNA gene, which has controlled the amplification of further ID elements. In rat, the formation of most ID genes has taken place in the last 3 million years (myr) (**Figure 4C**) [36, 37].

Rhesus macaque

Alu elements originated from a 7SL RNA shortly before the primate/rodent split. The precursor diverged into the B1 element in mouse, and into the Alu family in primates [32]. The current Alu element is dimeric and has a length of 280 nt. It retrotransposes via the L1 RT [35]. Alu amplification had a peak with the AluS subfamily around 45 million years ago (mya). Amplification rates are now declining, but several elements of the most recent AluY family that originated 25 mya are still active in the macaque line (**Figure 4D**) [39].

Human

The human Alu element is structurally close to the rhesus Alu element. As in rhesus, they are the most common SINE. Members of the AluY family are still active, although activity has declined more than 100x relative to primates 40-50 mya (**Figure 4E**) [32]. Nevertheless, with one new integration per 20 newborns, Alu elements belong to the most active TEs in the human genome [26].

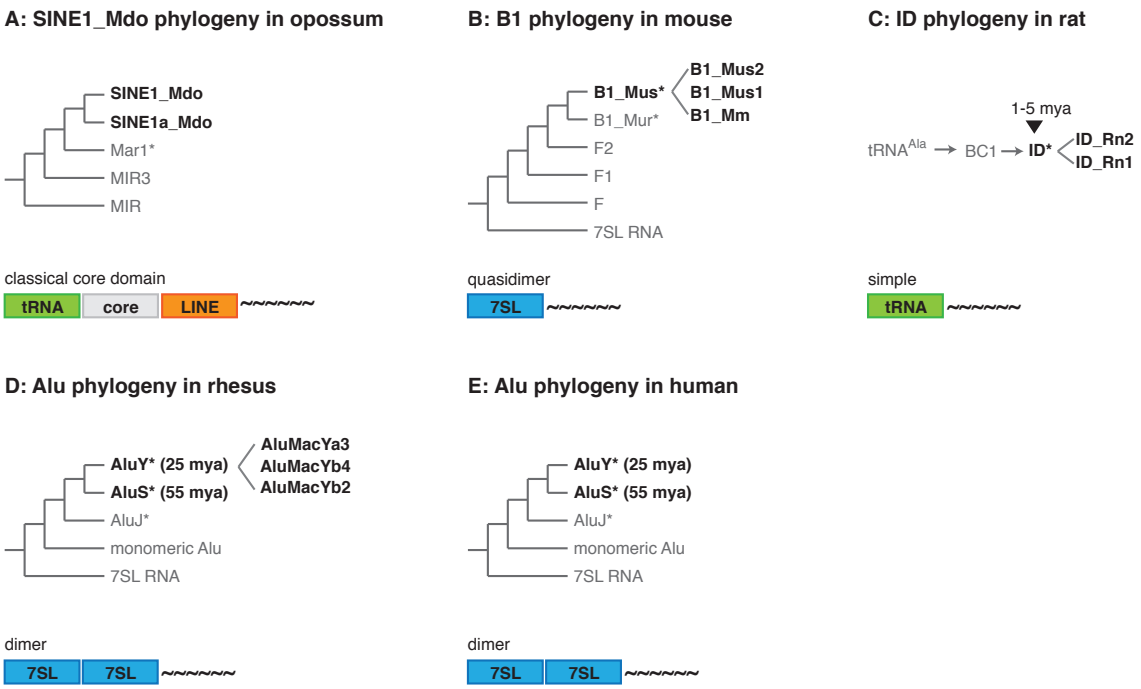


Figure 4: Phylogeny and structure of recently active SINE elements | A: SINE1_Mdo phylogeny in opossum. SINE1_Mdo and SINE1a_Mdo are the most recent, L1-mobilized TEs in opossum. Their structure is complex consisting of a head, core and tail domain. **B:** B1 phylogeny in mouse. The B1 element originated from the 7SL RNA. The family has undergone a strong diversification leading to the most recent B1_Mus* elements. B1 elements have a small region with internal duplications and are therefore classified as quasidimer. **C:** ID phylogeny in rat. The ID element originated from a tRNA. The non-coding BC1 gene is an ID element, which has played an important role in the amplification of the most recent ID TEs in the last 1-5 mry. **D:** Alu phylogeny in rhesus. Alu TEs originated from the 7SL RNA and are dimeric. The AluY family has undergone rhesus-specific amplifications. **E:** Alu phylogeny in human. As all primate Alu elements, Alu elements in human originated from a 7SL RNA, are dimeric and have developed human-specific elements *For all species, only selected SINE elements are presented in the phylogeny. The * indicates that a family consists of several TE members. The branch length does not reflect the phylogenetic distance. Structures are adapted from SINEbase [35].*

Active TE transposition can have negative effects on the host genome. New integration events that for instance interfere strongly with the structure or expression profile of a gene might reduce the fitness of the host. As a consequence, the host might have a reproductive disadvantage and the integration event will not be fixed in the genome. Nevertheless, mammalian genomes are repeat-rich and have allowed the integration and fixation of TEs. In the next chapter, I will therefore discuss the different forces that can influence integration and fixation of a TE in more detail.

1.2.3 Integration and fixation success of transposable elements

The distribution of transposable elements in the genome is shaped by a combinatory effect of integration and fixation biases. Integration describes merely the integration of a TE in a genomic region. Depending on the integration locus (e.g. in or outside a gene), the integration event can have different effects on the host, which influences whether the integration event remains and thus will be fixed, or not. It is still under debate whether LINE and SINE elements show integration biases, but it is widely accepted that they follow different fixation trends. Pavlicek *et al.* demonstrated that young LINEs and Alu elements are preferentially found in GC-poor genomic regions. In contrast, old Alu elements are enriched in GC-rich genomic regions, while old L1s accumulate in GC-low and gene-poor regions of the genome [30]. Levy *et al.* postulated that young TEs integrate more or less randomly [28]. This finding is supported by Ovchinnikov *et al.* who showed that young L1s are randomly distributed despite their AT-rich integration motif [29].

L1 elements are GC-poor. The human L1 consensus sequence has a GC content of 43.22%. The low GC content might explain why L1 elements are preferentially found in AT-rich regions in which they are less likely to disturb the local GC content. In addition, full-length L1s are large and integration in or close to a gene might easily lead to the disruption of the gene or some of its regulatory elements. As a consequence, L1s are fixed in gene-poor regions in which they are less harmful. SINE elements in contrast are GC-rich (e.g. AluY consensus: 63.12% GC, B1_Mus1 consensus: 57.82% GC). Despite their potential integration preference into GC-poor regions, they are more likely to be fixed in GC-rich regions corresponding to their own GC content. Close integration of two Alu elements in reverse-complement order is usually avoided as this might lead to local deletions or DNA recombination [41]. TEs can insert into each other, which allows the reconstruction of integration trajectories based on the assumption that evolutionary young elements will always integrate into evolutionary older elements [42]. Levy *et al.* could show that TE integration hotspots exist within the human genome. Alu elements for example, can integrate into the poly(A)-tail of already existing Alu elements. Importantly, old SINEs (MIRb, MIR3, AluJb, AluJo) seem to resist this integration. AluSx shows an in-sense integration bias into L1 elements. Interestingly, integration hotspots are more frequently found in gene proximity. A possible explanation is the low number of optimal (non-disturbing) integration sites in these regions, which forces TEs to integrate into the same locus [28]. Despite the fact that integration of a TE into a gene might be harmful, more than 90% of RefSeq genes contain TEs in their introns [43]. TEs, which integrated into introns, are underrepresented

close to the exon-intron boundary. Furthermore, SINE elements in mouse have an integration bias in antisense orientation to the gene, while human SINE elements are more likely to integrate in sense orientation. Alu elements possess a cryptic splice-site on their antisense strand, which might explain selection against antisense integration of Alus into human introns [43].

Mammalian genomes are rich in transposable elements, indicating that most of the observed TEs in our genome have little or no effect - either because the integration locus allowed for it, or because the genome developed defense mechanisms to reduce deleterious effects of TE integration. Although not harmful, the integration of a TE in close proximity or within a gene can contribute to the local noise levels of transcription and splicing.

1.2.4 Contribution of SINEs to transcriptional noise

Although some TEs can have positive influences on transcription and splicing by providing cis-regulatory elements, novel exons or epigenetic signals, the integration of most TEs is either neutral or deleterious. Approximately 5% of human alternatively spliced exons are derived from Alu elements [44]. Alu elements harbor cryptic donor and acceptor splice sites on their antisense strand [45]. If the Alu element integrates in antisense orientation to the gene, the spliceosome can recognize these cryptic splice sites, which leads to the exonization of the Alu element (**Figure 5A**). Because Alu elements are GC-rich, their integration can interfere strongly with the local intronic GC content. The spliceosome recognizes the GC amplitude between introns and exons, but integration of GC-rich SINE elements can disturb the pattern and lead to exon skipping (**Figure 5B**) [46]. Athanasiadis *et al.* provided evidence that two Alu elements that incorporate in reverse-complement order to each other in a gene, lead to back-looping of the pre-mRNA (**Figure 5C**). Interestingly, these backfolds (also called hairpin structures) are subject to RNA editing. The distance and divergence of backfolding Alu elements correlates negatively with the A-to-I editing frequency. However, editing does not interfere with the formation of backfolds, but instead increases the frequency of GC pairs that can be subject to methylation [47]. Gene body methylation is associated with increased gene expression, splicing enhancement and reduction of noise. Methylation of the Alu may thus be a counter mechanism to keep the gene active despite disruption by Alu integration (**Figure 5D**) [16, 48, 49].

DNA transcription and splicing are stochastic processes that can lead to a large variety of transcripts and expression profiles. Integration and fixation of transposable elements further contribute to this variety. Furthermore, stochastic influences are large if expression is low and when the gene

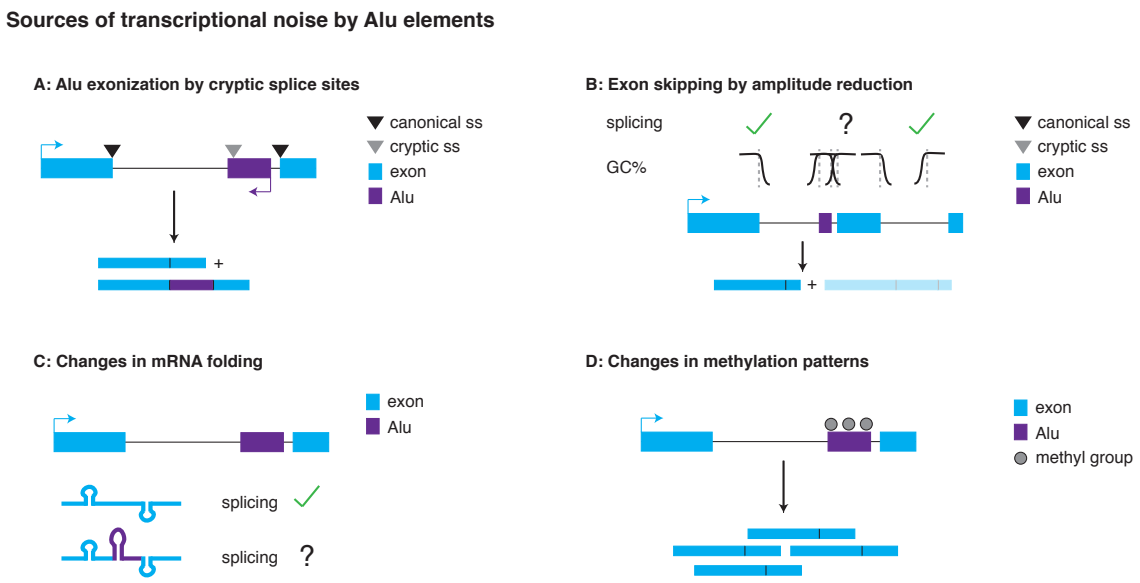


Figure 5: Transcriptional noise and Alu elements | A: Alu exonization. Alu elements harbor cryptic splice sites, which can lead to exonization of the Alu elements upon antisense integration into a gene. **B:** Exon skipping. Alu elements are GC-rich and their integration disturbs the local GC content leading to exon skipping [46]. **C:** mRNA folding. Integration can lead to mis-folding of the pre-mRNA/mRNA structure, which can interfere with splicing by hiding splice sites and other signals [47]. **D:** Methylation changes. Alu elements are subject to methylation, which can be enhanced by RNA editing. This leads to an increase in gene body methylation, which might enhance gene expression. *Abbreviations: ss = splice site.*

structure allows for many splicing reactions. Circular RNAs have been associated with all of the previously described properties of transcription and splicing: Low expression, alternative splicing and transposable elements. In order to understand whether they are functional elements of the mammalian RNA landscape, one therefore needs to understand their expression and splicing properties in the light of stochasticity.

1.3 CircRNAs

1.3.1 The re-discovery of circRNAs

Circular RNAs (circRNAs) were first described more than two decades ago, but the recent developments in large-scale sequencing technologies have thrown new attention on their properties and functions. CircRNAs are small and putative non-coding RNAs (100-1000 nt), in which the 3'-end and 5'-end of an RNA are covalently spliced to form a circular structure. They are known in at least four of the major kingdoms (Archaea, Bacteria, Plantae, Animalia). Their role is unknown, although functional cases in viral defense mechanisms (plants), miRNA binding and transcriptional

control have been described [50–52].

1.3.2 CircRNA biogenesis

The biogenesis of circRNAs is influenced by different factors in *cis* and *trans* [53]. Genes that are producing circRNAs (parental genes) are mainly protein-coding and are transcribed by RNA Pol II [54]. Promoters of parental genes are associated with higher levels of H3K27Ac (histone 3 acetylation at 27th lysine residue; used as a marker for active promoters) and a decrease in DNA methylation [55]. CircRNAs often overlap with linear transcripts of their parental gene and splicing is guided by the canonical splicing machinery [54, 56, 57]. Decrease of Pol II elongation capacity increases co-transcriptional splicing efficiency of linear transcripts [58]. In agreement with this finding, Ashwal-Fluss *et al.* observed that fruit flies carrying a slow-elongation mutation in Pol II had decreased levels of circRNAs. Furthermore, they could show that the increased length of circRNA flanking introns leads to a decrease in linear splicing efficiency [57]. Kramer *et al.* proposed that the length of circRNA flanking introns determines whether backsplicing occurs co- or post-transcriptionally. They further identified circRNAs, whose biogenesis is regulated by the combination of different hnRNPs (heterogeneous nuclear ribonucleoprotein) and SR (serine-arginine) family proteins (**Figure 6A**) [53]. The formation of circRNAs seems to rely on the presence of the canonical donor GT and acceptor AG splicing motif [54]. This finding is supported by Ashwal-Fluss *et al.*, who found a decrease of circularization in the fly *mbl* gene upon mutation of the canonical 5' splice site from GU to CA [57].

Interestingly, many research groups have observed an increased number of (small) repetitive elements that are in reverse-complement orientation (antisense) to each other in both circRNA-flanking introns. These elements facilitate backsplicing by the formation of hairpin structures that bring the circularizing exons close to each other (**Figure 6B**) [60–63]. In human for instance, many of the repetitive structures overlap with Alu elements [61, 64, 65]. Interference with the pairing capability by targeted mutagenesis of the Alu element or by RNA editing decreases circRNA levels [62, 66]. A similar enrichment of repetitive elements was also observed in mouse, pig and *C. elegans* [64, 67]. However, repetitive structures are absent in the flanking introns of *Drosophila* circRNAs, suggesting the existence of a repeat-independent mechanism for circRNA formation in this species [68]. Several groups have tried to analyze the repeat landscape of flanking introns in more detail, but have reached opposite conclusions. Veno *et al.* for instance found a positive correlation between the distance of a repeat to the backsplice site and the intron length. Consequently, they hypothesized

that repeats support backlooping only in small introns [67]. In contrast, Ivanov *et al.* reported a strong enrichment of repeats in the 1500 nt upstream and downstream of the circRNA splice site. Furthermore, they could show the repeat frequency can be used to predict circRNAs on a genome-wide level in human and *C. elegans* [64].

1.3.3 Alternative splicing of circRNAs

Certain loci produce multiple circRNAs that overlap partially with each other. These loci are called "hotspots" (**Figure 6C**). It is unknown why some genes have a higher potential of producing multiple circRNAs than others. The majority of hotspots was computationally predicted, but only a few groups have tried to validate circRNAs from the predicted loci [64, 67, 69]. Like linear transcripts, circRNAs can also be subject to alternative splicing. Exons are included or excluded and the acceptor and donor splice site can vary. These patterns of circRNA alternative splicing are tissue-specific and developmental stage-specific [70]. Zhang *et al.* showed that circRNAs contain novel exons. These exons were previously unknown, because of their low coverage in other studies, which were primarily

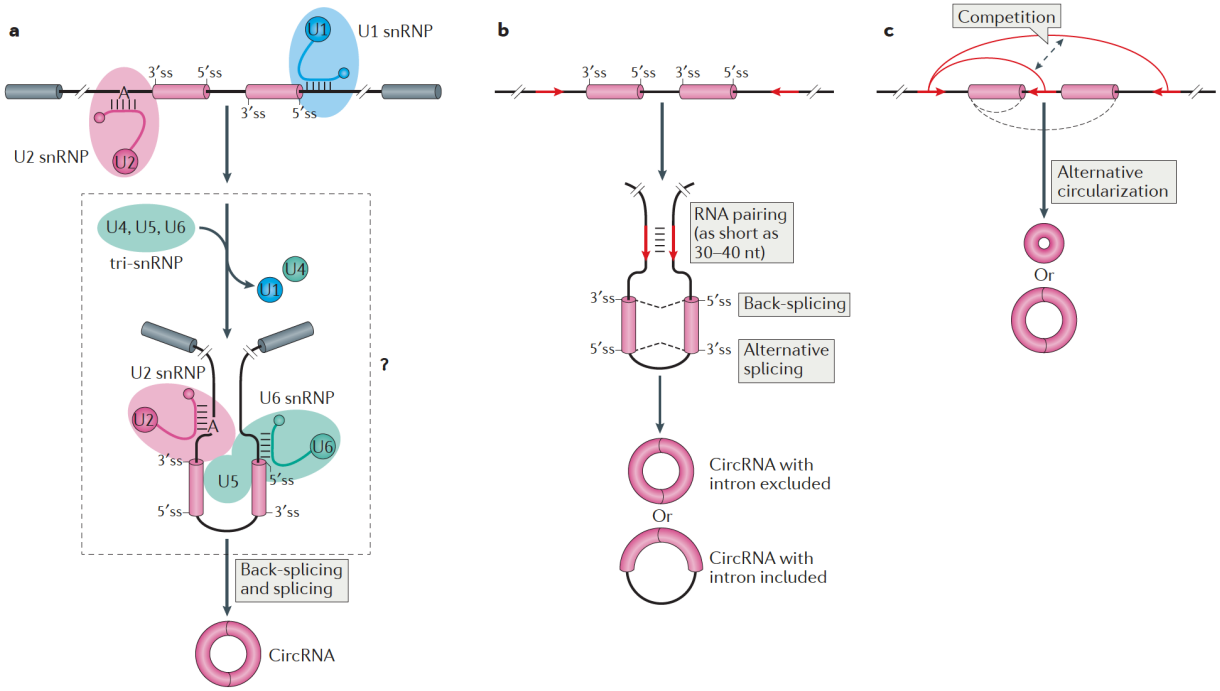


Figure 6: CircRNA biogenesis | A: CircRNAs are generated by a backsplicing mechanism, which is catalyzed by the spliceosome. With the potential help of hnRNPs, a 5'-donor site is joined to an upstream 3'-acceptor site. **B:** Repetitive sequences in the flanking introns can enhance circularization. **C:** A gene can produce several circRNAs. It remains unclear how these alternative circularization events are regulated. *Figure was adapted from Figure 1 in Chen et al. [59].*

looking at strongly expressed linear isoforms [71]. As part of their study Zhang *et al.* developed CIRCpedia - a database to collect alternative splicing patterns of circRNAs. However, it remains unclear how alternative splicing of circRNAs is regulated in the observed spatio-temporal manner.

1.3.4 CircRNAs - functional newcomers or by-product?

Little is known about how and to what extent circRNAs are part of the functional transcriptome. Based on a few characterized examples, following hypotheses for circRNA functions were proposed:

- Competing endogenous RNAs (ceRNAs).
- Alteration of gene expression by influencing transcription rates and alternative splicing.
- Molecular scaffolding to bind proteins.
- Transport of molecules to different cell compartments.
- Translation.

Competing endogenous RNAs

Currently, the best-described circRNA is ciRS-7 (also known as CDR1-as) from the *Cdr1* gene locus. ciRS-7 is highly abundant in mouse and human brain. It is strongly enriched ($>$ 70 sites) for binding sites of the miRNA miR-7. Both transcripts, ciRS-7 and miR-7, are co-expressed in murine neocortical and hippocampal neurons and ciRS-7 can efficiently bind and suppress miR-7 activity *in-vivo*. Ectopic expression of ciRS-7 in zebrafish, which has lost the *cdr1* locus, leads to a decrease of the conserved miR-7 and malformations in the developing, embryonic fish brain [72]. SNP density in potential miRNA binding sites is decreased in circRNA exons arguing for the idea that some circRNAs can be functional ceRNAs [73]. However, several studies were unable to find a strong correlation between predicted circRNAs and enrichment of miRNA binding sites [54, 74]. It is important to note that although circRNAs may function as ceRNAs, the majority may exhibit different or no function.

Alteration of gene expression

The drosophila gene *muscleblind*, a well-characterized splicing factor, produces a circRNA from its second exon. CircMbl is abundant in fly heads and MBL binding sites are enriched in the flanking introns of circMbl [57]. Binding of MBL promotes circularization at the cost of linear *mbl* (decrease

in production of *mbl*), an effect that was also observed with the linear and circular transcripts of the human *MBNL1* gene in HEK293 cells. Ashwal-Fluss *et al.* hypothesized that many circRNAs may, similarly to *mbl*, compete with a linear transcript for splice sites and thus enhance or interfere with transcriptional levels of linear mRNAs [57]. Quaking (QKI) is an RNA binding protein (RBP) and alternative splicing factor that acts as a homodimer. Knockdown of *QKI* by siRNA in mesHMLE cells leads to a decrease in abundance of more than 100 circRNAs (min. decrease 2-fold). QKI binding sites are enriched in the flanking introns of these circRNAs, and it is thought that binding of QKI to both flanking introns will bring them in close proximity by the formation of the QKI homodimer [75].

Molecular scaffolding

The *Foxo3* gene is a classical tumor suppressor. Du *et al.* showed that circFoxo3, a circular transcript of *Foxo3*, forms a ternary complex with the Ser/Thr protein kinase CDK2 and the cyclin-dependent kinase inhibitor p21. In the absence of p21, CDK2 activates cyclin E and promotes cell cycle progression. CircFoxo3 can bind p21 and CDK2. Du *et al.* suggested that the binding enhances the inhibitory potential of p21 on CDK2 by holding the two proteins in close proximity, thus inhibiting cell cycle progression [76].

Translation

Since the discovery of circRNAs, their role as a translational template for the synthesis of proteins has been debated. Several studies have shown that in *in-vitro* systems, circRNA expression plasmids have the potential to be translated [77, 78]. Yang *et al.* showed that circRNAs are associated with N^6-methyladenosine (m^6A), an RNA base modification that supports mRNA translation. Based on the enrichment of m^6A next to a circRNA, they identified 19 unique peptides likely originating from circRNAs [79]. The observation of translatable circRNA is supported by two additional studies: Pamudurti *et al.* found a small set of ribosome-associated circRNAs in different Drosophila cell lines. The peptide of circMbl was confirmed by northern blotting in the fly head [80]. Legnini *et al.* provided evidence for a peptide in human myoblasts originating from a circRNA in the *ZNF609* gene [81]. Nevertheless, the findings of these studies should be carefully evaluated as the number of detected peptides is small and functional evidence is missing. Similar to the circRNA transcript, the peptides could merely be a translational by-product.

Until now, only a handful of circRNAs is studied in detail. A slightly larger fraction is differentially expressed in human diseases, but the molecular link is unknown (see **Chapter 1.3.6**). It remains thus unknown, which fraction of circRNAs is truly functional or merely a by-product of stochasticity during transcription, splicing and translation of the parental gene.

1.3.5 Conservation of circRNAs

CircRNAs are frequently found in orthologous genes across different species. The observed overlap is used to support the hypothesis that circRNAs are functionally relevant. Veno *et al.* for instance, compared circRNAs between pig, mouse and human. Using the liftOver tool of the UCSC genome browser, they found that 88% of porcine circRNAs could also be detected in mouse. 20.4% of the splice sites used in mouse (fetal head) were identical to those used in pig [67]. In another study conducted by Rybak-Wolf *et al.*, approximately 28.5% of circRNAs shared the same splice site between human and mouse (allowing a 2 nt setoff) [66]. Contrarily, Zhang *et al.* proposed that circRNA landscapes are not conserved and undergo rapid evolutionary changes. Zhang *et al.* analyzed the conservation of repetitive elements in the flanking introns of several human circRNAs, but were unable to detect similar repetitive elements in species such as gorilla, rhesus macaque and mouse [61]. Until now, it remains a matter of debate whether circRNAs are conserved or subject to rapid turnover.

1.3.6 Biomedical relevance of circRNAs

The high number of different circRNAs has inspired researchers to investigate their involvement in human diseases. Several groups have shown that circRNAs are mis-regulated in different cancer types, such as pancreatic, colon or lung cancer [82, 83]. Interestingly, circRNAs tend to be down-regulated in cancerous tissues, which is associating them with tumor-suppressing roles. A handful of circRNAs is mis-regulated in Alzheimer disease [84], cardiovascular disorders [85] and osteoarthritis [86]. The exact contribution of circRNAs to these disease phenotypes still needs to be determined. In a first attempt to collect disease-related circRNAs to facilitate research, several groups have created databases for circRNA interactions with normal and disease-associated miRNAs, lncRNAs and proteins (Circ2Traits [87], CircInteractome [88], SomamiR 2.0 [89]). These databases will be helpful to filter the large quantity of predicted circRNAs for functionally relevant examples. CircRNAs have a very stable secondary structure. Recent studies have shown that circRNAs are detectable in human blood and serum samples and can be used as biomarkers for various diseases [86, 90,

91]. In addition, several groups have started to design circRNAs with sponge features to inhibit mis-regulated miRNAs [86]. Although associated with many human diseases, a direct molecular link between a circRNA and the observed disease phenotye is still missing. It is unclear whether the oberserved changes in gene expression and molecular phenotypes are due to the impact of the circRNA, its parental gene or a third source.

1.3.7 Detection and quantification of circRNAs by next-generation sequencing approaches

Detection of circRNAs

Identification of circRNAs relies currently on the identification of backsplice junctions (BSJs). The vast majority of the standard mapping tools is unable to detect these non-linear reads. Consequently, BSJ reads are discarded and labeled as "unmapped". In addition, circRNAs lack a poly(A)-tail and are therefore mainly present in rRNA-depleted (ribo-minus) total RNA pools. As a consequence, special remapping strategies are required to recycle BSJ reads at a significant level. In the general identification procedure, the n terminal 3'- and 5'-nucleotides of a sequencing read (referred to as anchors) are taken and are remapped independently on to the reference genome. Anchor pairs that map in a non-linear order on the same chromosome and strand are used as an indication for a BSJ if the extension procedure was successful (**Figure 7**).

The definition of successful depends on the detection tool and is influenced by the number of mismatches, the number of event-supporting reads and the presence of canonical splice sites. The number of read-supporting events is generally low (1-2 reads in 10 million reads), impeding the discrimination between noise and signal. The identification success of BSJ reads is additionally impaired by the number of multi-mapped reads due to low complexity DNA, (local) duplications,

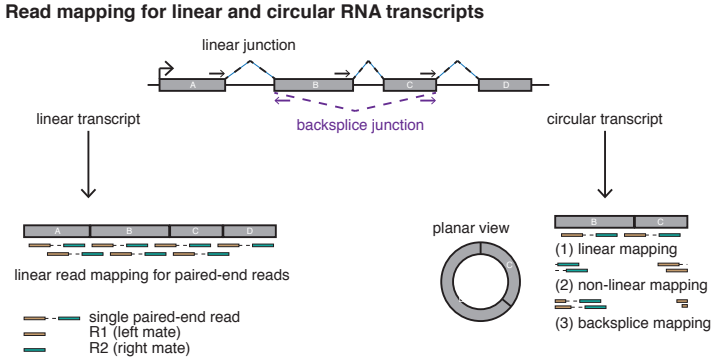


Figure 7: Read mapping for linear and circular RNA transcripts | Paired-end sequencing reads map in a linear order for linear transcripts (R1 upstream of R2). Reads that fall on a BSJ are characterized by non-linear mapping. Here, R2 is found upstream of R1. If the read maps directly on a BSJ, it needs to be split and the 3'-end of the read is found upstream of the 5'-end

Table 2: Performance of different circRNA identification tools | A: Tool structure and dependencies. For each tool, its language, mapper and dependencies are listed. **B:** Approximation of tool performance. Performance of each tool was categorized based on Hansen *et al.* [92], Gao *et al.* [96] and my own experience (marked with an asterisk *).

A:	Tool	de-novo	Language	Mapper	Dependencies
	CIRCexplorer2*	no	Python	Bowtie1/2	TopHat, bedtools, pysam, docopt
	CIRI*	yes	Perl	BWA	none
	circRNA_finder	yes	Perl	Star	samtools
	find_circ*	yes	Python	Bowtie2	pysam, samtools, bedtools
	KNIFE	no	Perl, R, python	Bowtie1/2	samtools, numpy, scypy, data.table
	MapSplice*	no	Python/C++	Bowtie1	samtools
	segemehl	yes	C, C	per se	samtools
	UROBORUS	no	Perl	Bowtie1/2	TopHat, samtools
B:	Tool	Run time	Memory	FDR	Sensitivity
	CIRCexplorer2*	medium	medium	low	high
	CIRI*	slow	high	high	high
	circRNA_finder	fast	medium	medium	low
	find_circ*	fast	low	medium	medium
	KNIFE	medium	high	medium	medium
	MapSplice*	slow	medium	low	high
	segemehl	medium	high	high	high
	UROBORUS	slow	low	low	low

rearrangements and paralogous genes in the reference genome. Since 2013, several groups have published strategies and algorithms to identify circRNAs from RNA-sequencing (RNA-seq) data. The first generation of tools relied merely on the identification of the BSJs, which was accompagnied by a high false-discovery rate (FDR), low sensitivity and little overlap between the different tools [92]. A comparison of five circRNA detection tools (find_circ [93], CIRCexplorer [61], circRNA_finder [68] CIRI [94], MapSplice [95]) by Hansen *et al.* showed only a moderate overlap of 16.8% between circRNAs identified by each of the five analyzed tools. To get an estimate for the false-positive discovery rate (FDR) of each algorithm, Hansen *et al.* ran each tool on normal and RNase R-treated RNA-seq data and calculated the number of predicted, but degraded circRNAs afterwards. Run time, memory requirements, FDR and sensitivity varied dramatically between the analyzed tools. Additionally, large differences were observed in the genomic distance between backsplice sites and the number of small circRNAs ($<$ 500 bp), which are probably false-positives. Hansen *et al.* concluded that a minimum of two detection tools is required to reduce the FDR to an acceptable level and to provide a solid data set for further work [92]. A summary of the different algorithms and their characteristics can be found in **Table 2**.

As circRNAs attract more and more attention, a second generation of tools is now under development, in which more statistical measurements are incorporated to discriminate between true and false BSJs. Gao *et al.* recently published a new version of CIRI (CIRI2), in which they incorporate

likelihood measures based on seed matching to judge the quality of BSJ reads. According to the authors, this change in the algorithm makes CIRI2 more reliable and faster than any of the other tools [96].

Szabo and Salzmann provide a comprehensive overview about the challenges of circRNA identification. They point out that criteria for circRNA assessment such as library types and preparation, count-based-expression quantification and cut-offs as well as RNase R enrichment scores are in need of common, more statistically based approaches to avoid large discrepancies between the findings of different studies [97].

Reconstruction of circRNA transcripts

All circRNA identification tools call circRNAs based on their unique BSJ, but they do not reconstruct the true exon-intron structure. CircExplorer2 circumvents the problem by making use of existing transcriptome annotations. It uses each annotated exon found between the backsplice sites to define the circRNA transcript [71]. It is evident that this assumption is not appropriate in many cases. For instance, Gao *et al.* showed that splicing patterns of circRNAs are as diverse as for linear RNAs. To reconstruct the circRNA structure, they used the differential enrichment of circRNA-contained junctions after RNase R treatment [70]. Metge *et al.* developed a reconstruction tool, which is based on long reads ($>$ 150 nt) from paired-end sequencing that map on the BSJ. Because of their length, long reads can provide information on the exon usage within the circRNA boundary [98]. However, while the approach of Gao *et al.* extracts additional information from the RNase R treated samples, the tool of Metge *et al.* is limited by the low number of usable reads.

Quantification of circRNAs

Currently, circRNA expression levels are determined by a simple, count-based approach (**C**ounts **p**er **M**illion = CPM). However, count-based approaches are the least accurate approaches and accuracy generally drops for weakly expressed transcripts [99]. Most circRNAs are identified based on a minimum number of 1-2 reads and statistical analysis needs to be perfomed carefully in the light of technical errors and transcriptional background [97]. In addition, proper normalization methods within and across libraries as well as replicates are underused in the circRNA research field. This can lead to strong under-estimations or over-estimations in circRNA abundances [97]. The importance of a circRNA is often judged by comparing the ratio of BSJ reads and linear reads at the BSJ. However, the ratio is influenced by the quantification method used and may bias results if transcripts of the

parental gene do not contain exons from the circRNA. Li *et al.* were the first who developed a tool for the parallel quantification of circRNAs and linear transcripts. In their pipeline, circRNAs are linearized and added to the total number of observed transcripts. Quantification is then performed with a modified version of *sailfish* (*sailfish-cir*). Interestingly, they were able to show that both the mis-quantification of circRNAs and the ignorance of its presence lead to mis-quantifications in the expression levels of parental genes [100].

Expression levels of circRNAs are low, therefore making it challenging to distinguish them from the transcriptional noise. Technical noise added by different detection and quantification pipelines further complicates this distinction and should be avoided by the usage of replicates and statistical methods [97].

1.4 Aim and focus of the thesis

CircRNAs constitute a newly discovered class of RNAs in the mammalian transcriptional landscape. They often exhibit low expression levels, have a stable secondary structure and their biogenesis is associated with repetitive structures in the flanking introns. Although the field of circRNA research has dramatically advanced in the last years, many questions remain to be (re)-addressed. For instance, it is still unclear how the combination of different transposable elements in the circRNA flanking introns influences their expression. Furthermore, it is debated to what extent circRNAs are conserved across different species.

Here, I am presenting a comprehensive data set of RNA-seq data for three organs of five species (opossum, mouse, rat, rhesus macaque, human) representing three mammalian lineages (marsupials, rodents, primates) to re-evaluate previously discussed questions and findings of circRNA characteristics. The number of circRNAs varies highly between detection pipelines and, as a consequence, many conclusions on circRNA size, expression levels, tissue specificity and conservation are different between datasets.

My first goal is therefore the development of a circRNA detection pipeline that 1) can successfully differentiate circRNAs from technical and biological artifacts and 2) can be used across different mammalian species and tissues. **Chapters 2.1-2.3** will provide details on the study design, the generation of the dataset for RNA sequencing and the different parts of the detection and quantification pipeline developed. The annotation of the exact exon-intron structure is challenging and I will show how I have used samples enriched in circRNAs to develop splicing graphs that can predict the circRNA structure.

I will then continue in **Chapter 2.4** to describe the structural properties of circRNAs across different species. Furthermore, I will introduce the concept of circRNA hotspots. I will demonstrate that they harbor important information that helps us understand how and why circRNAs are produced (**Chapter 2.4.3** and **2.8**). Following in **Chapter 2.5**, I will classify circRNAs based on their overlap between species. The here-presented dataset covers different mammalian lineages, which allows me to address the question if shared circRNAs evolved from a common ancestral circRNA or independently of each other.

By using linear regression models, I will assess several structural and functional properties of parental genes to provide evidence that circRNAs are preferentially found in a common genomic context created by low GC content, long introns and a high number of transposable elements (**Chapter 2.6**). In **Chapter 2.7**, I will show that transposable elements found in flanking introns are species-specific and still active suggesting that circRNAs did not evolve from a common ancestral circRNA, but developed independently in different species. I will show that the expression levels of a circRNA correlate positively with the abundance and amplification rates of TEs in their flanking introns, which further supports the idea of independent processes leading to shared circRNAs (**Chapter 2.8**).

Finally, I will discuss all results in light of current developments in the circRNA research field and argue for the hypothesis that circRNAs are a by-product of stochastic transcription, splicing and TE integration that is restricted to a specific genomic context (**Chapter 3**).

2 Results

2.1 Study design

In this PhD project, I analyzed the biogenesis and the evolutionary history of circRNAs. To interpret the latter correctly, I focussed on the following five mammals: *Monodelphis domestica* (opossum), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Macaca mulatta* (rhesus macaque) and *Homo sapiens* (human). Human and rhesus macaque are representatives for the primate lineage, mouse and rat represent the rodent lineage and opossum was chosen as outgroup to represent the marsupials. Primates and rodents are eutherian (placental) mammals, while opossum belongs to the metatherian (non-placental) clade of mammals. The split between the eutherian and metatherian lineage occurred 180 mya. Primates and rodents separated 90 mya, the common ancestor of mouse and rat lived 12-25 mya and the common ancestor of the rhesus macaque and human was found 25 mya (**Figure 8A**) [38, 101, 102]. The selection of these fives species allowed me to study not only species-specific circRNAs, but also the occurrence of lineage-specific (primate, rodent) and shared circRNAs (eutherian, therian). To capture different evolutionary dynamics, liver, cerebellum and testis were chosen as study organs. Samples from real tissues have the advantage that they reflect the *in-vivo* situation in greater detail than cell lines and cell line-specific batch effects are avoided. For each organ and species, I analyzed three distinct individuals with approximately the same age and sex (males only) to reduce the sample-specific signals as much as possible. An overview of the different samples is provided in **Supplementary Table 1**.

CircRNAs are characterized by low expression levels. However, they can be enriched by treatment with RNase R - an exoribonuclease that degrades linear transcripts, but is unable to act on transcripts with circular structure [103]. CircRNA identification relies strongly on the identification of BSJs. To gain sufficient information on the area around the BSJ, we therefore generated paired-end RNA-seq data for ribosome-depleted, total RNA of untreated and RNase R treated samples.

2.2 Library preparation and sequencing

**Division of tasks: All libraries were prepared with the help of Peggy Janich in the laboratory of David Gatfield at the Center of Integrative Genomics of the University of Lausanne.*

RNA-seq data was generated (median of $\sim$ 63 Mio single Illumina HiSeq reads per species) for total RNA from liver, cerebellum and testis for all five species (three males per tissue and species with exception of human liver). To biochemically enrich for circRNAs, a second data set was gen-

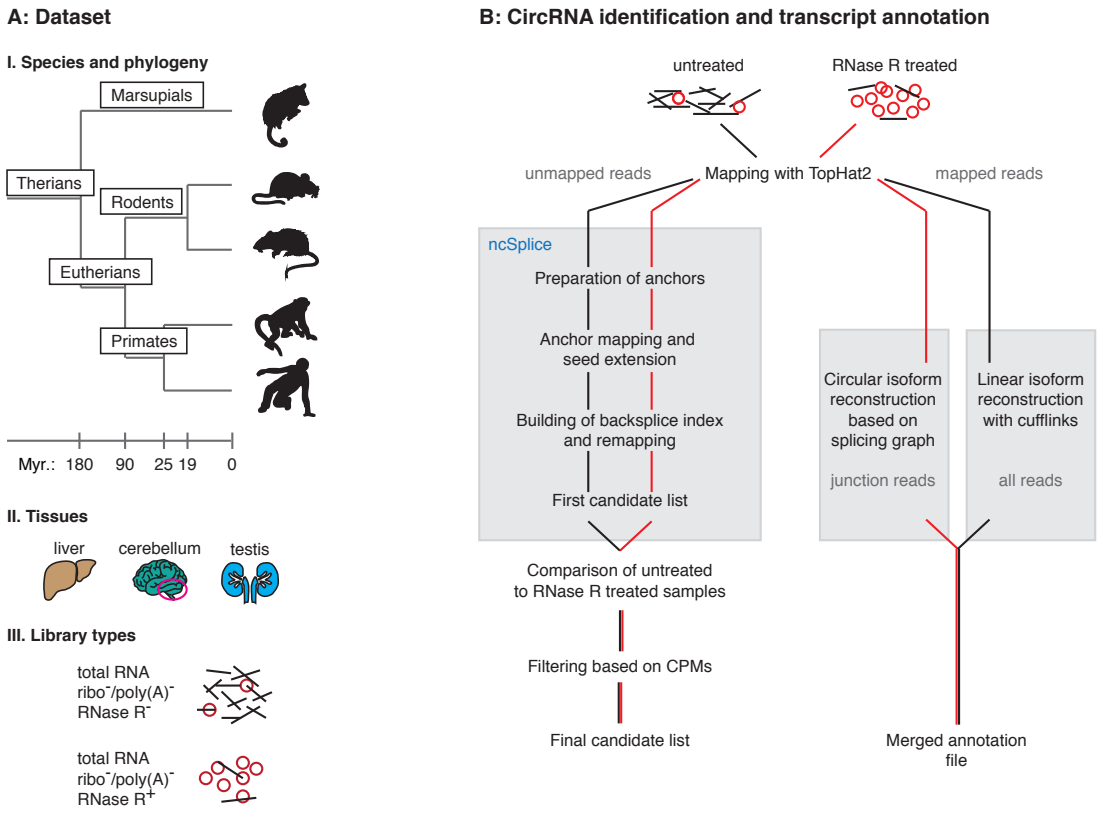


Figure 8: Overview of the dataset and the reconstruction pipeline | A: Dataset. CircRNAs were identified in five mammalian species (opossum, mouse, rat, rhesus, human) and three organs (liver, cerebellum, testis). For each sample, rRNA-depleted (ribominus) next-generation sequencing libraries were generated from untreated and RNase R treated total RNA. **B:** CircRNA identification and transcript reconstruction. Unmapped reads from RNA-seq data were remapped and analyzed with a custom pipeline. The reconstruction of circRNA transcripts was based on the junction enrichment after RNase R treatment. Further details on the pipeline are provided in Chapter 4.3.

erated for the same samples treated with RNase R prior to the library preparation (median of $\sim$ 58 Mio single Illumina HiSeq reads per species) (**Figure 8A**). RNase R treated libraries are different from the untreated libraries in the following characteristics: 1) The number of unmapped reads is higher in RNase R treated samples, 2) there is a 4.4-fold enrichment of multi-mapped reads upon RNase R treatment and 3) the number of BSJ reads in RNase R treated samples is 7.6-fold higher than the number of junctions reads in the untreated libraries. (**Supplementary Figure 1**). All observations are consistent with RNase R treated libraries being enriched in transcripts protected from degradation and transcribed, repetitive structures that can complicate the mapping process.

2.3 Identification and quantification of circRNAs

2.3.1 Development of a circRNA detection pipeline

I created a custom pipeline for the detection of circRNAs. The available detection tools in 2013/2014 (find_circ [93], MapSplice [95]) did not incorporate paired-end sequencing information, although the additional information given by the mapping coordinates of both read fragments decreases the FDR for circRNAs and other transcripts. In addition, MapSplice relied on the presence of an annotation file and find_circ only detected BSJs with canonical splice site motifs. Previous studies suggested that circRNA biogenesis is coupled to the presence of canonical splice sites, but I did not want to put a bias towards this hypothesis [54, 56, 57]. In short, the custom pipeline takes unmapped reads in bam-format to create a fastq-file with the 20 terminal nucleotides from the 5'- and 3'- end of each read (anchor pairs). Anchor pairs are independently mapped to the reference genome. In the next step, anchor pairs are filtered for being mapped on the same chromosome, the same strand and within a maximum distance of 100 kb to each other. Anchor pairs with a successful seed extension are then used to build a backsplice index for remapping of all unmapped reads. The remapping of reads allows to recover reads that span the BSJ with less than 20 bp and thus, to increase the read coverage per BSJ. The final candidate list is received after remapping and filtering for paired-end read pairs (**Figure 8B**).

In a collaborative project with Serghei Mangul (Eskin laboratory, UC Los Angeles), the custom pipeline was further adapted to detect fusion transcripts of genes lying on the same chromosome and trans-splicing events between different chromosomes. It was named ncSplice (**n**on-**c**o-linear **splic**ing). All parts of ncSplice that were used for this project are available on GitHub (`https://github.com/Frenzchen/ncSplice_circRNAdetection`). A detailed description of the pipeline is provided in **Chapter 4.3**.

2.3.2 Detection of circRNAs with ncSplice

I detected a total of 76,739 possible BSJs in opossum untreated samples, 67,249 in mouse, 72,855 in rat, 100,270 in rhesus macaque and 68,400 in human. Of those, 79-85% are unique to only one replicate and therefore, are likely to represent technical or biological artifacts (**Supplementary Table 2**). To filter the putative BSJs from noise, I compared untreated and treated libraries for BSJs that are present in both conditions and are enriched upon RNase R treatment (**Figure 8B**). I normalized the number of reads per junction (coverage) based on **c**ounts **p**er **m**illion (CPM) and kept only those

candidates with a CPM of at least 0.05 (corresponding to $\sim$ 1 BSJ read per 20 million reads). Some of the BSJs cluster around the same genomic locus with just a couple of base pairs difference in the start and stop coordinates. After manual inspection, it was evident that these reads are probably explained by mapping difficulties due to repetitiveness of the underlying genomic locus and were omitted. Upon these filtering steps, I received a final candidate list consisting of 1535 circRNAs in opossum, 1484 in mouse, 2038 in rat, 3300 in rhesus macaque and 4491 circRNAs in human (**Figure 9A; Supplementary Table 3**). In human, the number of circRNAs is overestimated due to the lack of two biological replicates for liver. In opossum, the total number of circRNAs in cerebellum is underestimated due to the low sequencing depth.

In human, rhesus and mouse, cerebellum possesses the highest number of circRNAs, followed by testis and liver (**Figure 9A**). However, the raw numbers are difficult to compare between species and tissues, because they strongly depend on the library depth and the mapping success. Therefore, I used the Shannon diversity index (H) to normalize circRNA numbers within each species and tissue. In ecology, the Shannon diversity index is often used for the quantification of the species diversity in a habitat [104]. The number of different species and their relative abundance within the population are used to calculate richness and evenness of the observed population. H is not only a good measure for the species composition of a population, but can also assess the diversity (richness) and the relative expression levels (evenness) of a transcriptome [105]. In the here-discussed case, all circRNAs in a given tissue (habitat) belong to one population. Individual circRNAs are

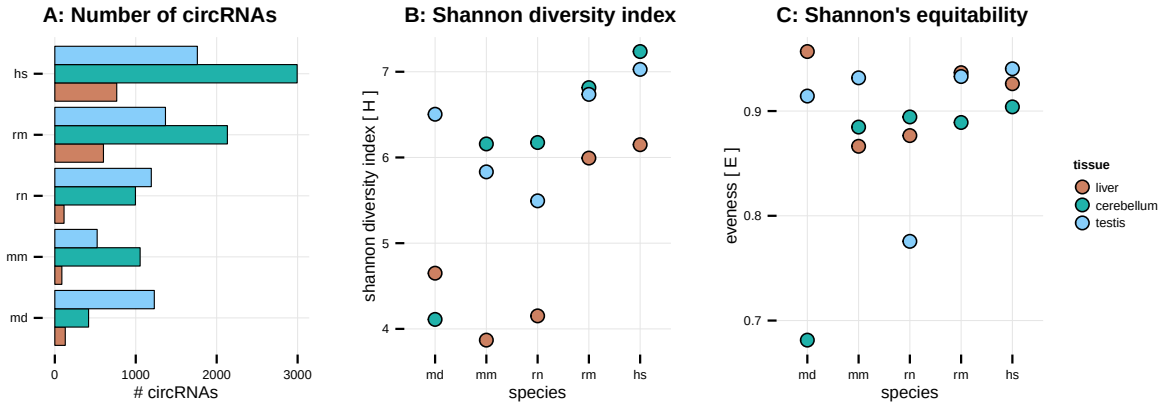


Figure 9: CircRNA frequencies | A: Total number of circRNAs. The total number of detected circRNAs is plotted for each species in liver (brown), cerebellum (green) and testis (blue). **B:** Shannon diversity index. H for circRNAs in different tissues and species is plotted. **C:** Shannon's equitability. EH for circRNAs in different tissues and species is plotted. *Abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human.*

treated as individual species from the population that can be common (high expression) or rare (low expression). To understand if the expression profiles of the detected circRNAs are different when compared to each other or evenly distributed, I calculated *H* for each circRNA population in a given tissue. Cerebellum is the tissue with the highest *H*, followed by testis and liver. The low *H* in opossum cerebellum is likely due to the low library depth of this tissue (**Figure 9B**). *EH* - Shannon's equitability - can assess the evenness of expression levels in the transcriptome. It takes a value between 0 and 1, with values equal to 1 meaning that frequencies are evenly distributed. For transcriptome data, a value of 1 corresponds to equal expression levels of all observed transcripts. For circRNAs, *EH* is close to 1 for most of the tissues indicating that many circRNAs have similar (low) expression levels. Cerebellum and liver have a lower *EH*, pointing to some circRNAs that occur more frequently than other circRNAs in these tissues (**Figure 9C**).

In summary, the total number of annotated circRNAs is highest in cerebellum, followed by testis and liver. However, the population structure is similar in all three tissues: There are many weakly expressed circRNAs and only a few more highly expressed transcripts.

2.3.3 Annotation of circRNAs

The detection of a BSJ provides only little, although important information on a circRNA. CircRNAs have low expression levels and it is difficult to assess to what extent exons in between the BSJ belong to the circRNA or to an overlapping, often more highly expressed linear transcript. To reconstruct the exon-intron structure of the circRNA isoform, I used the RNase R treated libraries, because exonic junctions, which are part of the circRNA should also be increased after RNase R treatment. They can be used to represent the circRNA transcript structure (**Figure 10A**). I took all junction reads that are either outside (outer junctions) or inside (inner junctions) the BSJ in a given gene. In RNase R treated samples, outer junctions are expected to have low coverage due to the treatment. Therefore reads from these junctions were taken to reconstruct a background distribution for the coverage decrease. I then compared the coverage of each inner junction to the background distribution and defined inner junctions, for which the coverage was outside of the 90%-confidence interval (CI) of the background distribution, as circRNA transcript junction. Observed differences in coverage between inner junctions of the same gene, might be due to 1) the presence of several circRNA isoforms or 2) overlapping circRNA transcripts. To identify the most strongly expressed isoform and to distinguish between different circRNA transcripts, I constructed a circRNA splicing graph for each gene and calculated the most abundant isoform based on the mean junction coverage

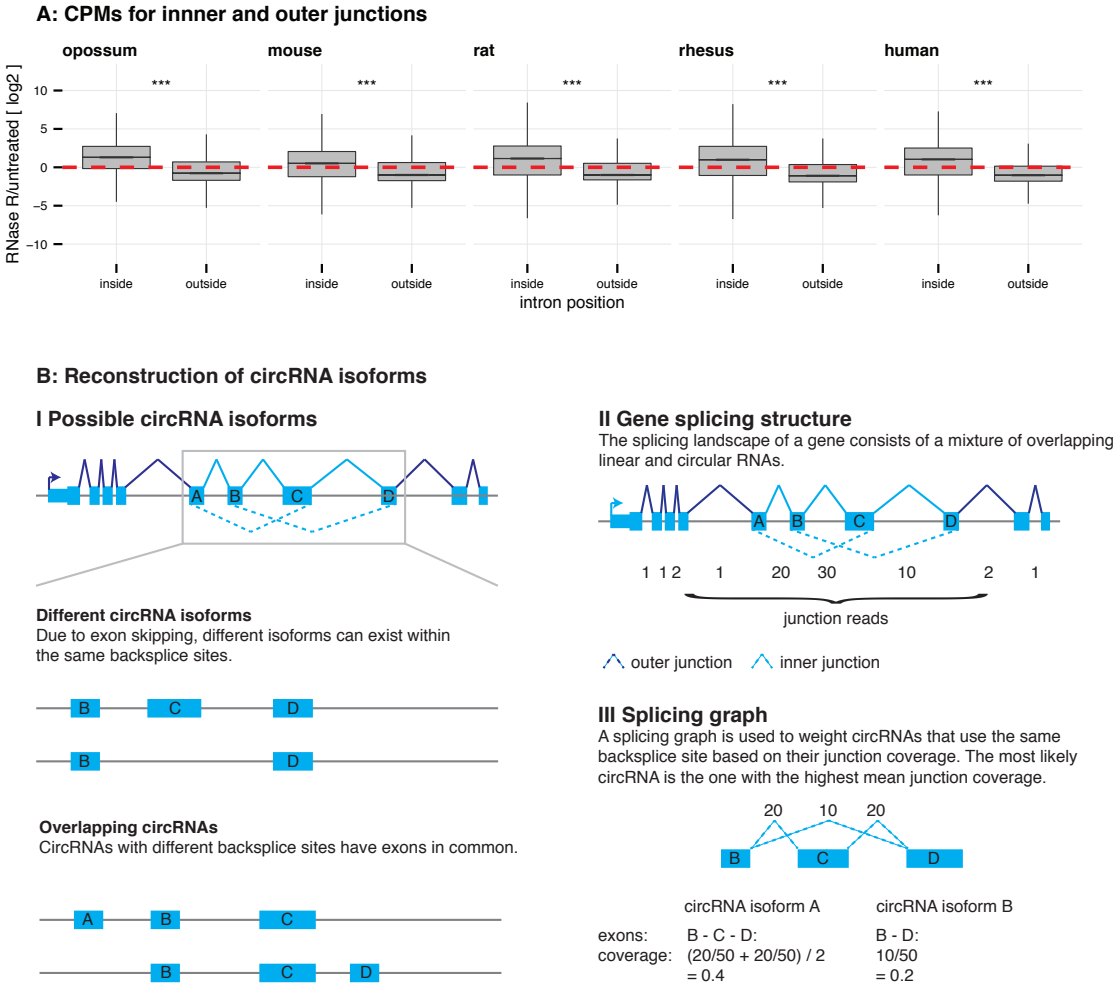


Figure 10: CircRNA transcript reconstruction | A: CPMs for inner and outer junctions. Reads from all libraries of a given species were pooled and significance assessed with a one-tailed Student's t-Test (for all comparisons, p-value < 0.001). **B:** Reconstruction of circRNA isoforms. CircRNAs can have different transcript structures (I). To reconstruct structures, junctions were grouped into inner and outer junctions based on their position to the circRNA (II). The most likely transcript was calculated using splicing graphs (III). *For panel A, outliers were removed for plotting.*

(**Figure 10B**).

2.4 CircRNA properties

2.4.1 General properties of circRNAs

CircRNAs differ in several features from other non-coding RNAs: They are small (median genomic length 11,000 bp, transcript size between 400-800 bp, median exon number 3) (**Figure 11A-C**). In addition, circRNAs overlap mainly with protein-coding genes and they are tissue-specific (**Figure 11D-E**).

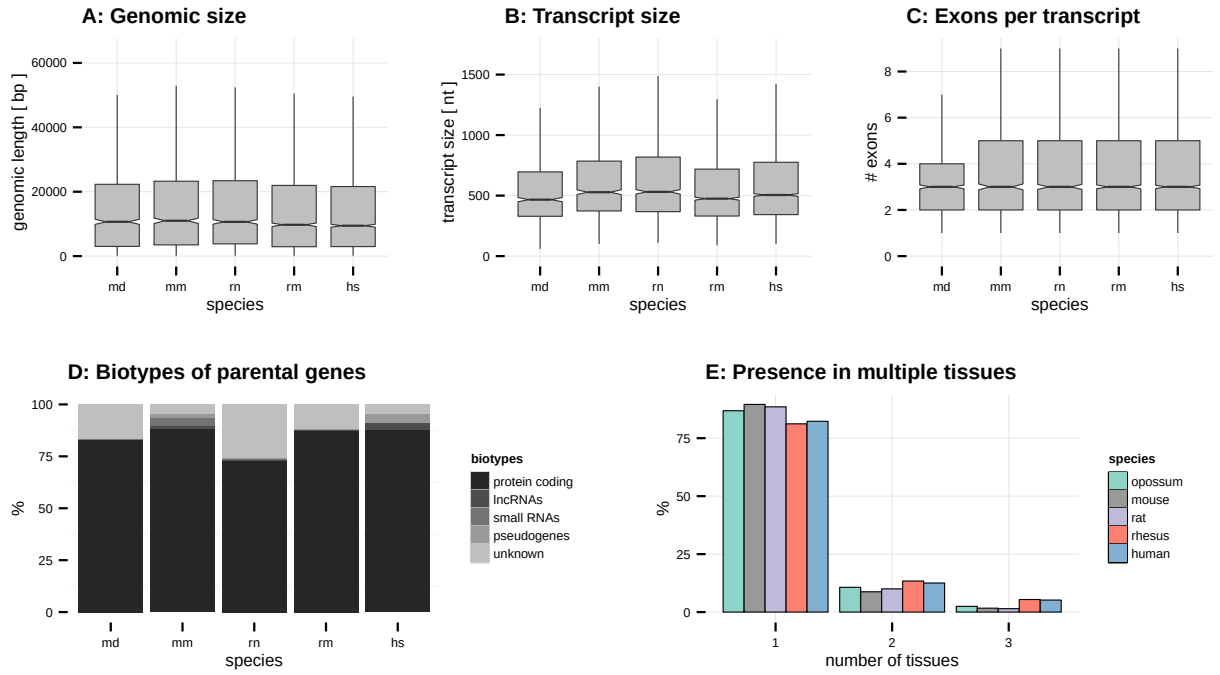


Figure 11: General properties of circRNAs | A: Genomic size. The genomic size (bp) of circRNAs is plotted for all species. **B:** Transcript size. The transcript size (nt) of circRNAs is plotted for all species. **C:** Exons per transcript. The number of exons in circRNAs is plotted for all species. **D:** Biotypes of parental genes. For each species, the frequency (%) of different biotypes in the circRNA parental genes was assessed using the ensembl annotation. CircRNA loci that were not found in the ensembl annotation were marked as "unknown". **E:** Presence in multiple tissues. For each species, the frequency (%) of circRNAs detected in one, two or three tissues is plotted. *For panel A-C, outliers are not plotted. Abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human.*

2.4.2 Validations

**Division of tasks: qPCRs were performed under supervision of Peggy Janich in the laboratory of David Gatfield at the Center of Integrative Genomics of the University of Lausanne.*

We confirmed a small subset of circRNAs for mouse, rat and human in liver, cerebellum and testis

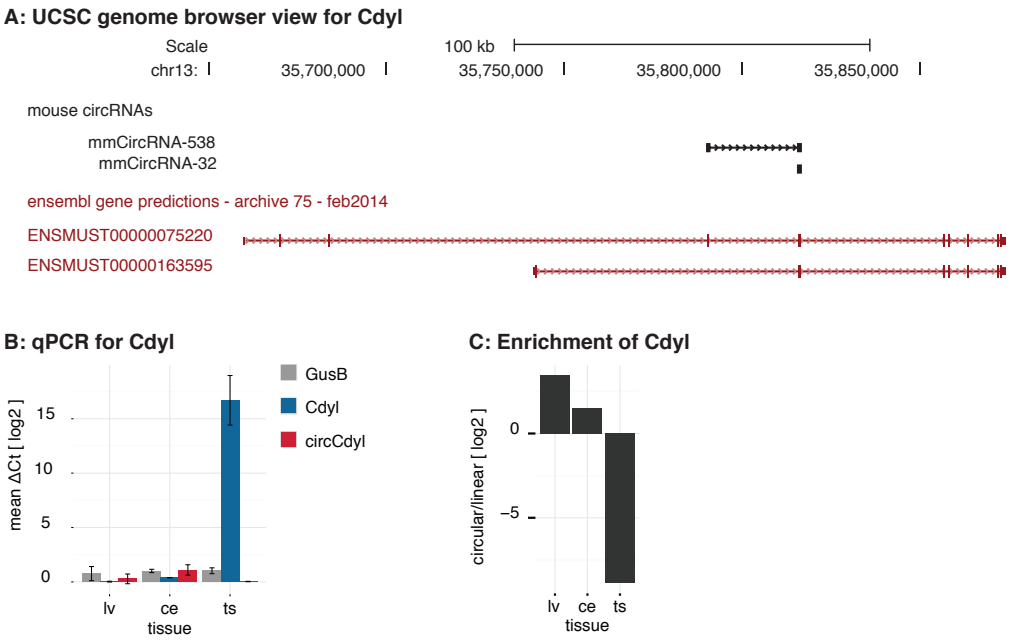


Figure 12: Genomic locus and validation of circCdyl | A: UCSC genome browser view for Cdyl. Depicted are the circRNA reconstructions (purple) and the ensembl reference transcripts (red) for the Cdyl locus in the mouse genome. **B:** qPCR for Cdyl. Mean ΔCt for *Cdyl* and circCdyl in liver, cerebellum and testis. Values were normalized to the housekeeping gene *GusB*. **C:** Enrichment of Cdyl. The log2-ratio of ΔCt values between circCdyl and *Cdyl* in different tissues is plotted. *Abbreviations: lv = liver, ce = cerebellum, ts = testis*

via *quantitative real-time PCR* (qPCR) with primer specific for the BSJ. We selected circRNAs based on 1) strong expression in a given tissue, 2) strong differences in expression between tissues and 3) presence in several species. 23 murine circRNAs were tested in liver, three in cerebellum and two in testis of which we were able to confirm 93% (26/28) (**Supplementary Table 4**). The *Cdyl* gene (*chromodomain protein Y-Like*) produces a circRNA from its 7th exon (circCdyl). My analysis predicted circCdyl expression to be high in cerebellum, low in liver and absent in testis. It also predicted a strong tissue specificity for *Cdyl* in testis and low expression in cerebellum and liver. The qPCR results agreed with all predictions that were made for the expression profiles of circCdyl and *Cdyl* (**Figure 12**).

2.4.3 CircRNA hotspots

Parental genes can give rise to multiple often overlapping circRNAs. These genes are known as circRNA hotspots, but the frequency of and the mechanism behind this phenomenon has not been studied. I decided to analyze circRNA hotspots in more detail, as the presence of a hotspot might

indicate the presence of a strong and therefore detectable signal leading to the biogenesis of circRNAs. In the subsequent chapters, I am defining a circRNA hotspot as any parental gene that produces at least two genomically overlapping circRNAs independent of their expression levels or the tissue they are found in. To understand the properties of circRNA hotspots in more detail, I used different CPM thresholds (0.01-0.1 CPM) to analyze the changes in the total number of detected circRNAs and the number of circRNAs per gene in respect to different CPM thresholds. With a decreasing CPM, the number of circRNAs varies between 6292-17,281 at 0.01 CPM and 682-2187 at 0.1 CPM. While at 0.1 CPM only 33-48% of circRNAs are in a hotspot, the number increases to 59-76% at 0.01 CPM (**Figure 13A**). The exact numbers for each species are provided in **Table 3**. The median number of circRNAs in a hotspots increases from 2-3 circRNAs at 0.1 CPM to 5-6 circRNAs at 0.01 CPM (**Figure 13B**). Hotspots are furthermore characterized by the presence of a dominant isoform that in average exhibits at least 50% of the summed expression levels in a hotspot. Transcripts within a hotspot overlap more often in their middle-exons than with the first or last exon. A hotspot can produce different circRNAs in multiple tissues (**Figure 13C-F**). **Figure 13F** provides an example for a circRNA hotspot in the *Crebrf* gene. The gene produces nine different circRNAs with different structure, tissue preference and expression levels. mmCircRNA-366 is the dominant circRNA in this hotspot contributing 33.5% to the observed summed expression.

Table 3: CircRNA frequency as function of CPM threshold | Table provides an overview of the total number of circRNAs that were detected in a given hotspot at 0.1 and 0.01 CPM (column 2 and 4, corresponding percentages are indicated in brackets behind the total number of circRNAs) and the total number of hotspots (column 3 and 5) at 0.1 and 0.01 CPM in each species.

Species	#CircRNAs at 0.1 (%)	#Hotspots at 0.1	#CircRNAs at 0.01 (%)	#Hotspots at 0.01
Opossum	682 (33)	87	6292 (59)	1184
Mouse	707 (36)	96	6660 (65)	1232
Rat	1016 (47)	151	7711 (67)	1457
Rhesus	1607 (44)	273	14,029 (73)	2804
Human	2187 (48)	408	17,281 (76)	3353

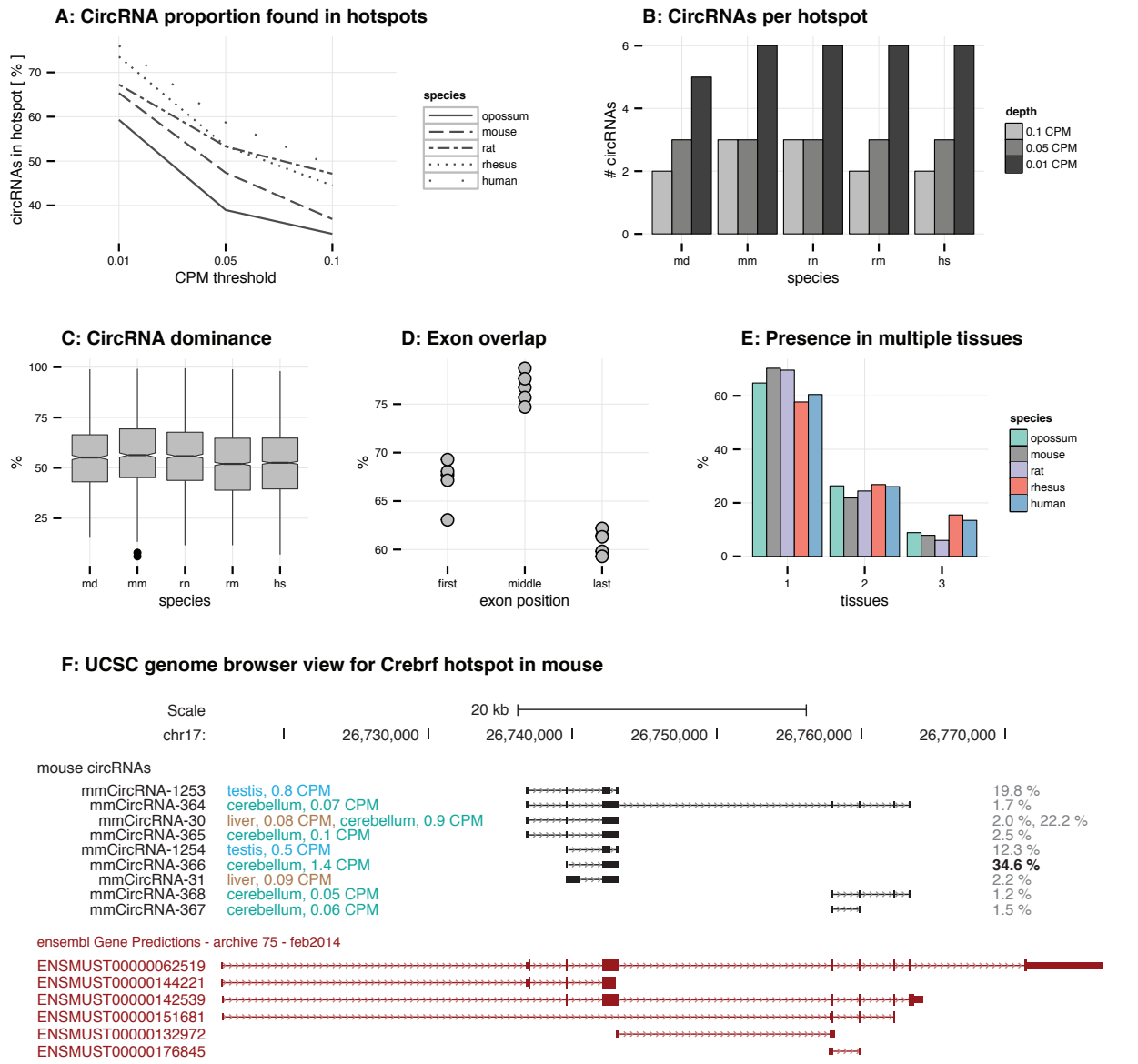


Figure 13: Hotspot properties | A: CircRNA proportion found in hotspots. For each species, the fraction (%) of circRNAs that is found in a hotspot at 0.01, 0.05 and 0.1 CPM is plotted. **B:** CircRNAs per hotspot. Each bars represents the median number of circRNAs found per hotspot and species at different CPM thresholds. **C:** CircRNA dominance. The CPM of each hotspot was defined as the sum of all circRNA CPMs from within the hotspot. The contribution of the most highly expressed circRNA was defined as $CPM_{strongest\ circRNA}/CPM_{hotspot}$. **D:** Exon overlap. For each hotspot, the fraction (%) of circRNAs overlapping with the first (=acceptor), middle and last (=donor) exon is calculated. **E:** Presence in multiple tissues. For each species, the frequency (%) of hotspots detected in one, two or three tissues is plotted. **F:** UCSC genome browser view for Crebrf hotspot in mouse. Depicted are the circRNA reconstructions (black) and the ensembl reference transcripts (red) for the Crebrf locus. CircRNAs from this hotspot are expressed in different tissues as underlined by the color of each transcript (liver = brown, cerebellum = green, testis = blue). CPM values are indicated next to the tissue. The contribution (%) of each circRNA to the summed hotspot expression was added to the right of each transcript. mmCircRNA-366 is the strongest circRNA contributing 33.4% of expression to the hotspot. *Abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human.*

2.5 CircRNA overlap between species

2.5.1 Frequency of shared circRNAs

Many studies have identified circRNAs that were produced from orthologous genes and subsequently, claimed that circRNAs are highly conserved [66, 67]. However, none of the studies was primarily designed to gain insights into the circRNA structure and abundance across different species and therefore, a suitable dataset to support this claim is missing. The here-presented dataset of circRNAs across different mammalian species and lineages allows to gain deeper insight into the subset of species-specific circRNAs. Furthermore, potentially orthologous circRNAs can be defined and their evolutionary history investigated. For the following analyses, I distinguished shared circRNAs based on five groups:

- *Therian circRNAs:* The circRNA is present in opossum and in at least three out of four species.
- *Eutherian circRNAs:* The circRNA is not present in opossum, but in at least three out of four species.
- *Rodent circRNAs:* The circRNA is present in mouse and rat (referred to as lineage-specific).
- *Primate circRNAs:* The circRNA is present in rhesus macaque and human (referred to as lineage-specific).
- *Species-specific circRNAs:* The circRNA is present in only one out of five species, or can not be grouped in any of the other four groups.

Furthermore, I decided to define the structural overlap between circRNAs based on three different levels (**Figure 14A**):

1. Overlap based on common parental genes (referred to as level 1).
2. Overlap based on common loci within parental genes (referred to as level 2).
3. Overlap based on the first and last exon of the circRNA (referred to as level 3).

The level-1 classification defines two circRNAs as overlapping as soon as they have orthologous parental genes. Under this assumption, 4681 distinct clusters, of which 5% (236) are shared between

Table 4: Expected and observed frequencies of parental gene clusters | The expected value was calculated as $N * \sum p(i)$ with *p(i)* representing the probability of an individual event (e.g. having an one-to-one ortholog in the set of opossum parental genes) and N is the number of trials (observed number of events in the different groups).

Group	Therian	Eutherian	Rodents	Primates
Expected	50	117	250	652
Observed	236	299	304	790

all five species, were identified. 80.0% of clusters are species-specific (opossum: 14.7%, mouse: 10.6%, rat: 12.0%, rhesus: 16.7%, human: 26.7%). The identified number of clusters in the individual groups (therian, eutherian, rodents, primates) is significantly larger than the expected overlap (Pearson's chi-squared test, p-value < 0.001). For instance, based on the one-to-one orthology of parental genes from the five species used, one would expect to find 50 therian clusters if they were randomly sampled. However, 236 clusters exist between the five species (**Table 4, Table 5**).

In the level-2 classification, circRNA coordinates within each gene are collapsed to identify the maximal locus from which circRNAs can be produced. Collapsed loci are synonymous to the previously described hotspots. By applying this criterion, a total of 5429 distinct clusters were found. Similar to level-1, 4.8% of clusters are annotated as therian and 75.5% are species-specific. Level-3 classification defines overlapping circRNAs based on the usage of the same first and last exons in different species. 10,064 distinct clusters were identified in total, of which only 1% represent therian and 85.6% species-specific clusters (**Table 5, Figure 14B**). The number of identified clusters increases with the classification stringency. However, the proportion of different orthology groups is similar between the level-1 and level-2 classification. In contrast, level-3 is characterized by an increase in species-specific circRNAs. The stringency increases strongly from the level-1/level-2 classification to level-3 explaining the increase of species-specific circRNAs.

Table 5: Overview of detected clusters under different classifications | Table summarizes the frequency (%) of different clusters based on the circRNA locus classification and overlap between species. The total number of detected clusters in each group is indicated in brackets.

Group	Level-1 %	Level-2	Level-3
Therian	5.0 (236)	4.8 (260)	1.0 (104)
Eutherian	6.4 (299)	6.8 (369)	2.6 (263)
Rodents	0.9 (43)	1.9 (102)	1.9 (192)
Primates	7.6 (356)	10.9 (594)	8.9 (893)
Opossum	14.7 (688)	15.9 (862)	14.2 (1418)
Mouse	10.6 (498)	8.1 (438)	9.9 (989)
Rat	12.0 (563)	12.5 (681)	14.4 (1438)
Rhesus	16.7 (782)	15.3 (833)	18.3 (1835)
Human	26.1 (1226)	23.7 (1289)	28.8 (2887)
Total	4691 cluster	5428 cluster	10,064 cluster

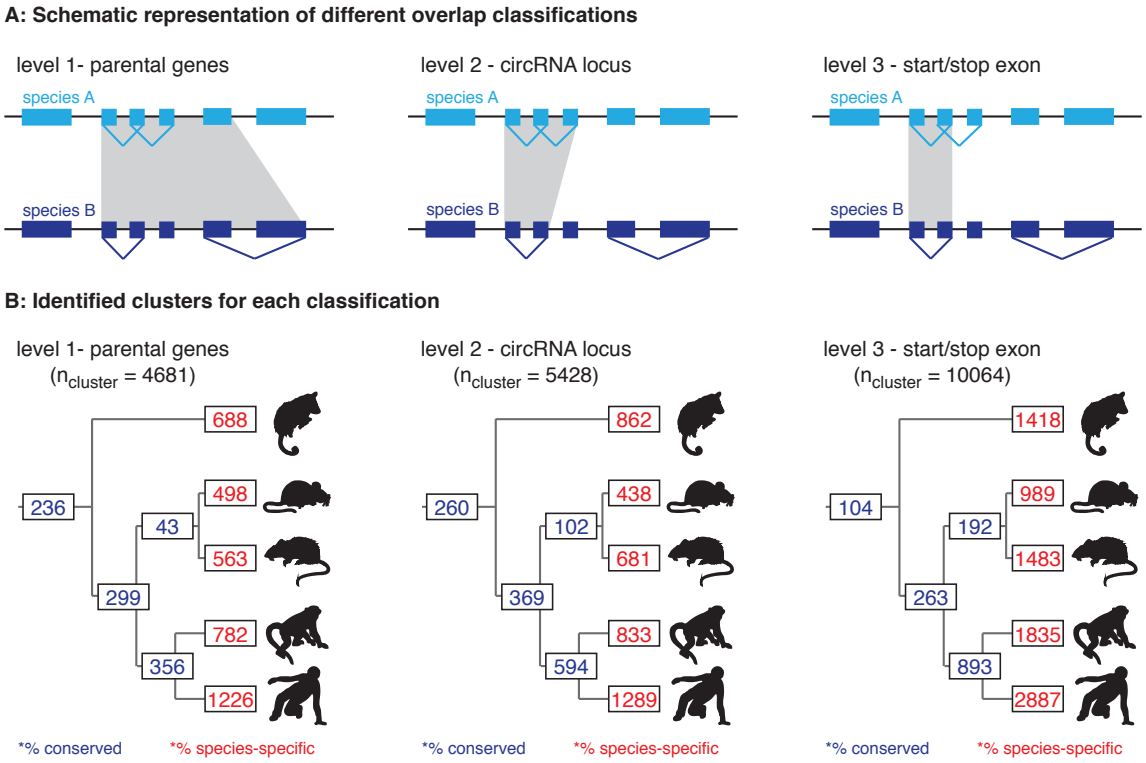


Figure 14: CircRNA overlap between species | A: Schematic representation of different overlap classifications. The overlap can be defined on the gene level, based on the locus within the gene or the same first and last exons. **B:** Identified clusters for each classification. The differences in the total number of identified clusters are caused by the different classification criteria. The occurrence of "0" in the level-1 classification is probably caused by the usage of 1:1-orthologous genes that had to be present in all five species analyzed. If not a 1:1-ortholog in the five species, they were automatically classified as species-specific.

I decided to continue with shared circRNAs from the level-2 classification for the subsequent analyses, because it is less stringent than level-3, but provides a better resolution for circRNAs that are shared between eutherians, rodents and primates. In addition, level-2 reflects the previously defined hotspots. Shared hotspots may indicate the presence of a strong signal for the production of circRNAs, which itself might be lost if shared circRNAs are analyzed based on the parental gene (level 1) or the exact first and last exon (level 3).

2.5.2 Level 2 circRNAs and their properties

PhastCons scores are conservation scores based on multiple alignments and known phylogenies that describe the conservation levels of a given base pair [106]. Increased conservation scores for any genomic element (e.g. an exon, UTR, regulatory sequence) can indicate that the element itself is functional and therefore, has been kept without large changes in the genome. To understand whether

circRNA exons distinguish themselves via higher conservation scores from the remaining exons of the parental gene, I calculated the conservation scores separately for each exon in the parental gene and compared the different exon types with each other (exons inside and outside the circRNA). PhastCons scores were available from the UCSC genome browser server for mouse, rat and human. Interestingly, the ratio of phastCons scores between exons that are inside and outside the circRNA was significantly larger than expected (one-tailed Wilcoxon rank sum test with log2-transformed ratios and an expected value μ=0, for all comparisons: $p < 0.001$) (**Figure 15A**).

Species-specific genes and transcripts can exhibit lower expression levels than genes and transcripts that are shared between multiple species. Therefore, I decided to compare expression levels between therian circRNAs and species-specific circRNAs. To allow comparison between the different species without relying on the CPM values, I calculated the ratio between BSJ reads and spliced reads that were covering the direct flanking introns (circ-ratio). If the locus contained several circRNAs, I first calculated the ratio for individual circRNAs and then, summed the ratio of circRNAs within one locus for further comparison. For all species, the circ-ratio is larger for therian than for species-specific circRNAs (**Figure 15B**). However, the effect is less distinct (but still significant) for the mean and median circ-ratios (data not shown). The observed differences in using the summed, mean or median circ-ratio may indicate that the number of circRNAs is higher in shared circRNA loci than in species-specific circRNA loci and is thus driving the observed differences. Shared circRNA loci are therefore more often circRNA hotspots.

The expression levels of protein-coding genes are well conserved across homologous tissues of different species. LncRNAs in contrast, exhibit a species-specific clustering with low correlation between expression levels [22, 23]. If circRNAs were a conserved property of the genome, then circ-ratios should be similar across tissues and species. To investigate this hypothesis, I calculated pairwise rank-correlations (Spearmann's rho) between the circ-ratio for all tissues and species. Independently of whether the summed, mean or median circ-ratio is used, circRNA clustering is driven by a mixture of tissue-specific and lineage-specific signals: They cluster first by lineage (rodents, primates), but within each lineage, homologous tissues are closer. In contrast, parental genes cluster mainly by species (**Figure 15C**). This might be surprising, however, clustering was performed with a small subset of genes representing a specific subset of genes (therian circRNA parental genes).

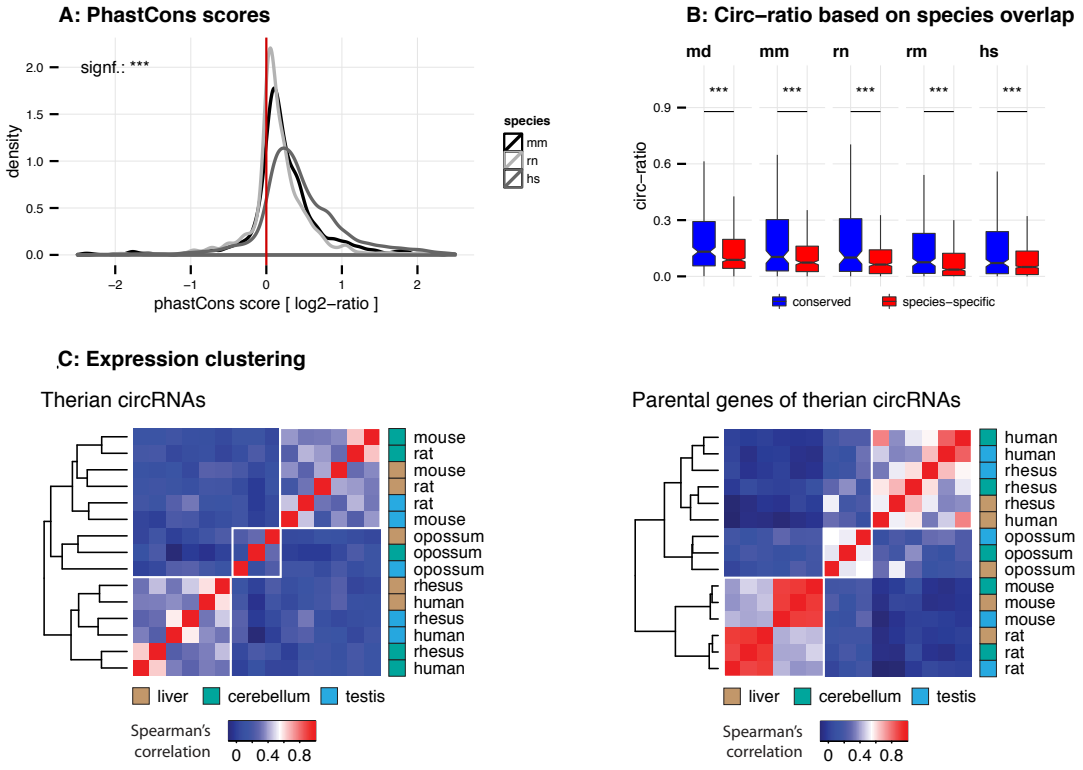


Figure 15: Properties of level 2 circRNAs | A: PhastCons scores. PhastCons scores were calculated for each exon using the conservation files provided from ensembl. For each gene, the mean phastCons scores for exons inside the circRNA were compared to the mean phastCons scores for exons outside the circRNA. The distribution of the obtained ratios is plotted. Significance was assessed by a one-tailed Wilcoxon rank sum test using log2-transformed ratios and an expected value of μ=0 (for all comparisons: $p < 0.001$) **B:** Circ-ratio. Circ-ratio was calculated with BSJ reads and spliced reads that were covering the direct flanking introns. Significance was assessed by a one-tailed Mann-Whitney U test using the summed circ-ratio of shared (blue) and species-specific (red) circRNAs. **C.** Expression clustering. Hierarchical clustering of circ-ratio for therian circRNAs (n=260) and FPKMs (fragments per million mapped reads) for parental genes of therian circRNAs (n=251). Heatmap presents Spearman's rank correlation coefficients between pairs of samples. Sample clustering is further depicted by squares on the left of each plot representing the different organs (liver = brown, cerebellum = green, testis = blue). *Abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human.*

2.5.3 Divergent and parallel evolution

The phylogenetic relationship between different species is reconstructed by combining phenotypic and genomic traits. If the traits originated in a common ancestor, they are defined as homologous between species (**Figure 16A**). Starting from the common ancestor, they can evolve in different directions - a process, which is also known as divergent evolution. However, similar environmental conditions can lead to the parallel evolution of analogous traits between species without common origin (**Figure 16B**). One famous example of parallel evolution for instance, is the evolution of

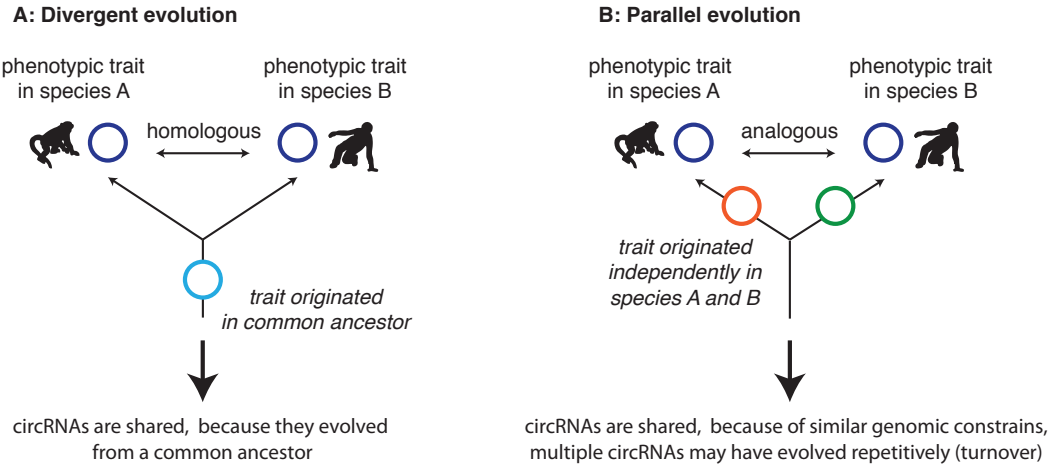


Figure 16: Schematic representation of the concepts of divergent and parallel evolution | A: Divergent evolution. The circRNA was present in the common ancestor of species A and B. **B:** Parallel evolution. The observed circRNAs in species A and B developed independently due to common genomic constrains. Turnover over of circRNAs (e.g. change of the backsplice site) with time is possible.

a camera-type eye in cephalopodes and vertebrates. In both groups, the need of seeing led to the development of the same eye structure [107]. The emergence of similar coloration patterns in different animals is a second example in which the need of interacting with the environment (e.g. mating display, communication with other individuals, protection from predators) has created similar coloration patters in different species (reviewed in [108]). The presence of a similar trait across species does therefore not always imply homology, and common ancestry can be a misleading assumption. In the previous chapters, I showed that circRNAs are shared between species. This overlap can be due to a common ancestral circRNA or instead, be based on structural or functional constrains, which are shared by their parental genes. In the latter, circRNAs, which are shared between species are not conserved and did not originate from a common ancestor. Instead, they evolved in parallel driven by a similar genomic environment provided by parental genes. To investigate this hypothesis, I analyzed different genomic components to understand if any of them are shared between circRNA parental genes.

2.6 CircRNA parental genes

2.6.1 Genomic and functional characteristics

To investigate whether the presence of a circRNA in multiple species is explained by divergent or by parallel evolution, I analyzed the different properties of parental and non-parental genes. I collected information on the genomic length, GC content and the number of exons and transcripts of each gene to approximate its structure. Furthermore, I scored its splicing properties based on the splice site amplitude (difference between intron and exon GC content at splice site) and incorporated functional aspects such as the expression levels of a gene, its phastCons scores and GO annotations. Last, I analyzed the repeat landscape in each gene. Previous research in human and mouse has shown that repetitive structures as well as DNA-binding proteins, which are present in the flanking introns of circRNAs, can support their biogenesis [64]. In human, these repetitive structures overlap often with Alu repeats [65, 75]. In species different than human, it is yet unknown whether there are specific repeat families that influence circRNA biogenesis.

GC content

To facilitate the classification of parental genes, I divided coding genes into distinct groups based on the GC content of the five isochores: L1 with $< 37\%$, L2 with 37-42%, H1 with 42-47%, H2 with 47-52% and H3 with $> 52\%$. In agreement with previous studies, coding genes in human and rhesus macaque have a bimodal GC distribution (gene enrichment in L2 and H3). Mouse and rat as representatives of the murideae are characterized by a uniform distribution with a gene peak in H1. In opossum, genes are enriched in GC-low isochores (L1-L2) [9]. After calculating the percentage of parental genes in each isochore, I found that parental genes are strongly enriched in GC-low isochores (**Figure 17**). In opossum, 94% of parental genes are found in L1 and L2. In mouse and rat, parental genes are mainly in L2 and H1 (89 and 82% respectively). In rhesus and human, parental genes peak in L2 followed by L1 and H1 (92% and 93%). Thus, low GC content is associated with the presence of a parental gene.

Exon and intron structure

For all coding genes, I categorized introns and exons according to the presence of a circRNA and their relative position to the circRNA. Exons were divided into 1) non-parental, 2) parental, but outside of the circRNA and 3) parental and inside the circRNA. Similar to exons, introns were

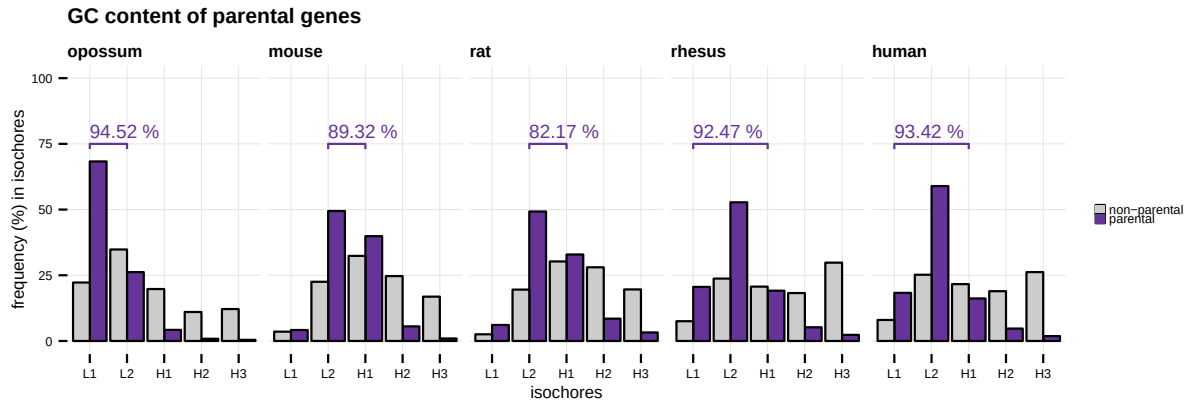


Figure 17: GC content of parental genes | Coding genes were classified into L1-H3 based on their GC content (grey). The percentage of parental genes was then calculated in each group (purple). The percentage of parental genes in L1-L2 (opossum, mouse, rat) and L1-H1 (human, rhesus macaque) is indicated above the respective graph (purple writing).

grouped into 1) non-parental introns, 2) parental, but outside of the circRNA, 3) parental and flanking to the circRNA and 4) parental and inside the circRNA. For each category, I calculated the median length and GC content. Exons that are contained in the circRNA have a significantly smaller GC content than other exons from the same parental genes (one-tailed and paired Mann-Whitney U test, p-values mostly < 0.001) (**Supplementary Table 5**). The exon length does not differ. CircRNA flanking introns are significantly larger than other introns inside and outside of the circRNA independent of the GC content (one-tailed Mann-Whitney U test, p-values < 0.001) (**Figure 18A and D**). The difference is more pronounced between circRNA flanking introns and non-parental introns (6.7-9.9x longer) than between circRNA-flanking introns and parental-outside introns (4.9-6.4x longer). Exon and intron structure are depending on the isochore, in which the gene is located [3, 109]. The observed differences between the analyzed intron and exon types are in agreement with the enrichment of parental genes in GC-low isochores.

Next, I calculated the GC amplitude at each exon boundary, which I defined as the difference between the GC content of the last 250 intronic bp and the first 50 exonic bp of each splice site. Amit *et al.* described a strong correlation between GC amplitude at the exon boundary and nucleosome occupancy in the exon creating a strong signal for the splicing machinery. In contrast, a low amplitude was correlated with exon skipping [3]. Consistently in all species, I detected an increased amplitude at splice sites of circRNA exons in comparison to parental-outside exons and non-parental exons (one-tailed Mann-Whitney U test, p-values < 0.001) (**Figure 18B, D**). The correlations are stronger for introns, which suggests that the amplitude increase is driven by a reduced GC content

in the intron (**Figure 18C-D, Supplementary Table 6**). In summary, circRNA flanking introns are long and GC-poor, circRNA exons have lower GC content and the GC amplitude at circRNA splice sites is larger than at other splice sites (**Figure 18D)**.

Complementarity and repeats

I analyzed the complementarity and repeat structure of coding genes. Similar to Ivanov *et al.*, I used megaBLAST to align all annotated coding genes with themselves to identify regions of complementarity in sense and antisense orientation of the gene [64, 110]. In general, self-complementarity correlates negatively with GC content (**Figure 19A**). Parental genes show a stronger level of self-complementarity in sense and antisense than non-parental genes from the same isochore. Furthermore, at least 2/3 of circRNAs and hotspots have at least one complement or reverse-complement alignment in their flanking intron. Next, I intersected introns of all genes with the RepeatMasker annotation. RepeatMasker provides information on the exact coordinates, repeat family and similarity to the consensus sequence of individual members of a TE family in a given species. For this analysis, I counted the number of repeats in each intron. I did not normalize counts by intron length, because it is unknown whether the overall frequency in an intron or the repeat enrichment at a specific position is of importance. In the latter, division by the intron length can complicate the interpretation of the results. However, I performed a similar analysis, in which I focused on length-matched introns instead of parental genes. This analysis is discussed in **Chapter 2.7**. Across all species, parental genes are 6.9-9.0x enriched in repeats (sense and antisense) than non-parental genes (**Figure 19B**). The enrichment frequency is similar for both orientations, but the total number of repeats is higher for antisense integrations. The high number of annotated repeats in the parental genes explains the higher level of self-complementarity.

Replication time, gene expression steady-state levels and haploinsufficiency

For human, I additionally analyzed functional data from three independent studies to assess parental genes for their replication time [111], the steady-state levels of gene expression [112] and haploinsufficiency scores [113]. The replication time refers to the order, in which chromosomal segments are duplicated during cell division. The replication time for instance correlates with evolutionary divergence and SNP density and can thus provide interesting information about the chromosomal locus [114, 115]. For coding genes, GC content correlates negatively with the replication time of a gene. Interestingly, parental genes are among the early replicating genes in each isochore (**Figure**

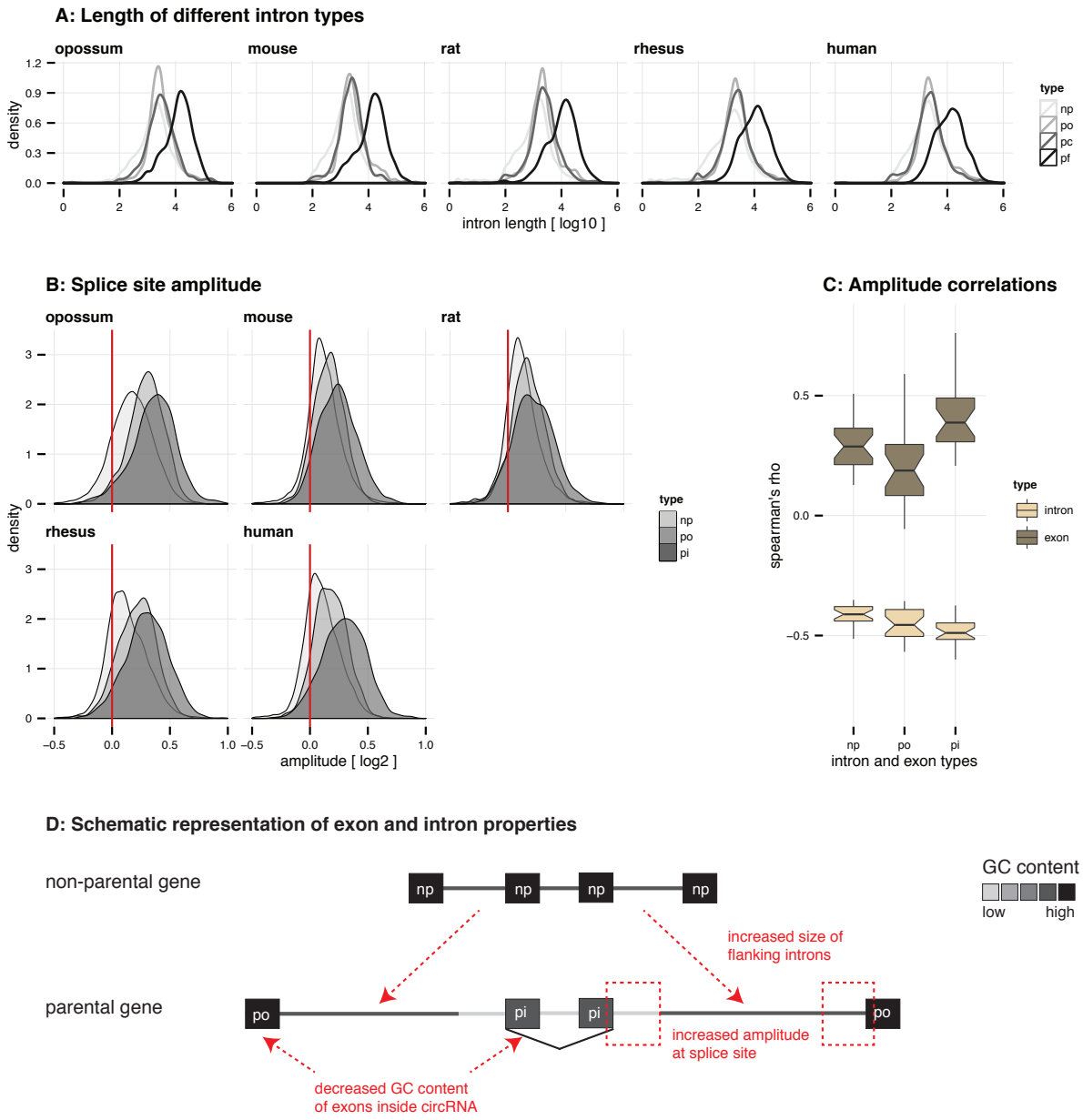


Figure 18: Properties of circRNA introns and exons | A: Length of different intron types. Distribution of median intron length (log10-transformed) is plotted for different intron types in each gene. **B:** Splice site amplitude. Distribution of median GC amplitude (log2-transformed) is plotted for different exon types. Red vertical bar indicates value at which exon and intron GC content would be equal. **C:** Amplitude correlations. Plotted is the correlation (Spearman's rho) between the amplitude and GC content of introns (light brown) and exons (dark brown). **D:** Schematic representation of exon and intron properties. *Abbreviations: np = non-parental, po = parental-outside of circRNA, pf = parental-flanking of circRNA, pi = parental-inside of circRNA.*

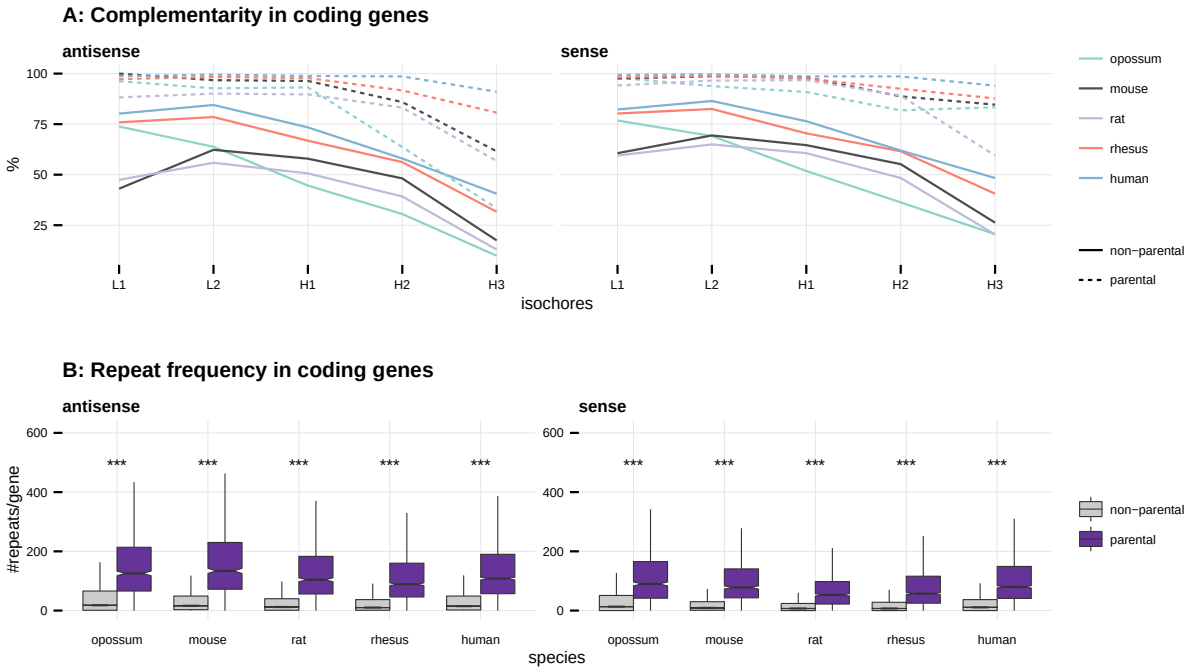


Figure 19: Complementarity and repeats | A: Complementarity in coding genes. Each coding gene was aligned to itself in sense and antisense orientation using megaBLAST. The proportion of each gene involved in an alignment was calculated and plotted according to its isochore. **B:** Repeat frequency in coding genes. The total number of repeats was counted in coding genes using the RepeatMasker annotation. Significance was calculated using a one-tailed Mann-Whitney U test. For all comparisons, the p-value is smaller than 0.001. *Outliers for panel B were removed prior plotting.*

20A).

The steady-state levels of gene expression are the results of mRNA transcription and mRNA decay. Pai *et al.* used microarrays to measure mRNA decay and steady-state expression levels across a time course of four hours in lymphoblastoid cell lines [112]. Based on their data, the mean steady-state expression level for parental genes is higher than for the remaining genes from the same isochore (**Figure 20B**). The decay rates for non-parental and parental genes do not differ strongly (data not shown), which suggests that the observed differences are mainly driven by enhanced transcription.

Haploinsufficiency describes the phenomenon, in which a diploid organism has only one functional allele of a gene, because the second allele is impaired by mutation(s). As a consequence, gene expression levels are reduced causing an altered phenotype. Haploinsufficiency is therefore a good measure to assess the dominance of an allele. Recently, Steinberg *et al.* developed a genome-wide haploinsufficiency score (GHIS) based on co-expression networks, *dN/dS* and single nucleotide polymorphisms (SNPs) for a large set of coding genes [113]. The GHIS of parental genes in human

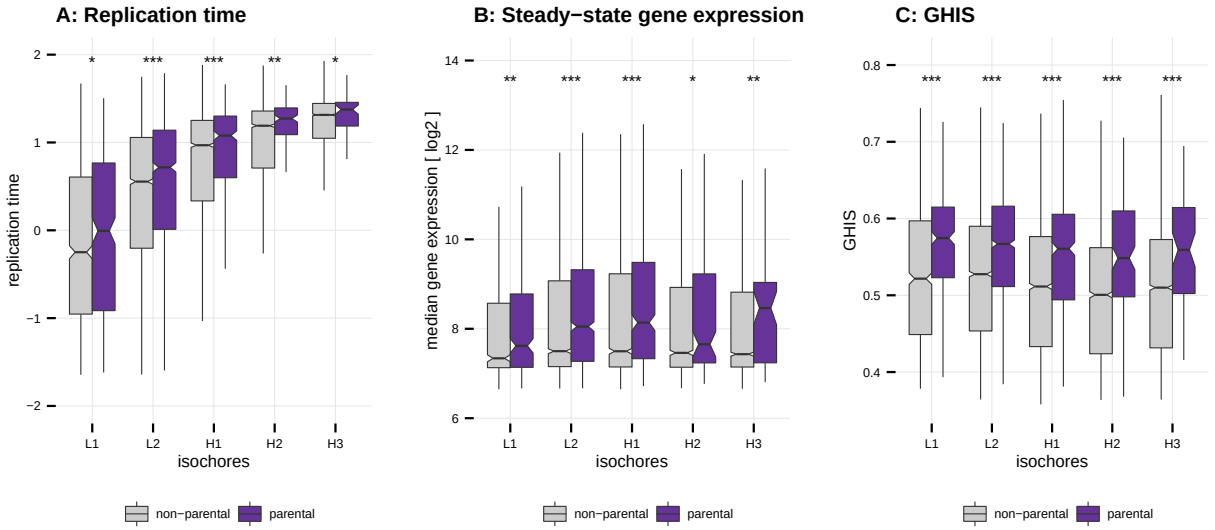


Figure 20: Functional properties of human parental genes | A: Replication time. Values for the replication time were used as provided in Koren *et al.* [111]. They are normalized to a mean of 0 and a standard deviation of 1. Difference between non-parental genes (n_{total}=18,134) and parental genes (n_{total}=2058) was assessed by a one-tailed Mann-Whitney U test. **B:** Gene expression steady-state levels. Mean steady-state expression levels were used as provided in Pai *et al.* [112]. Difference between non-parental genes (n_{total}=14,414) and parental genes (n_{total}=2058) was assessed by a one-tailed Mann-Whitney U test. **C:** GHIS. GHIS was used as provided in Steinberg *et al.* [113]. Difference between non-parental genes (n_{total}=17,438) and parental genes (n_{total}=1995) was assessed by a one-tailed Mann-Whitney U test. *Outliers for all panels were removed prior plotting. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.05.*

is higher than for other genes from the same isochore, indicating that the organism could be more sensitive to changes in the expression levels of parental genes (**Figure 20C**).

In summary, parental genes in human are early replicating in their GC environment, they have higher steady-state expression levels and a higher GHIS, suggesting that their transcriptional activity is important for the organism.

Gene ontology

I analyzed gene ontology (GO) annotations for parental genes in mouse, rat and human against a background dataset of expressed genes from isochores L1, L2 and H1 (GO annotations for rhesus macaque and human were not available). The background dataset was restricted to genes from these isochores, because they account for more than 90% of the observed circRNAs. The GO process category is enriched in terms directly connected to the tissue's function (metabolic processed for liver, neuronal process in cerebellum and spermatogenesis in testis). Interestingly, in all three tissues and species, functional categories are enriched in GO terms connected to binding, modification

and transferase- and kinase-activity. No consistent enrichment was found for the GO category of components genes are expressed in (data not shown).

2.6.2 Linear regression models to analyze parental genes

In the previous analyses, I showed that parental genes differ in structural (e.g. GC content, intron length, amplitude, repeat structure) and functional traits from non-parental genes (e.g. replication time, steady-state expression levels, GHIS). However, it is unclear which of the traits affect the presence of a circRNA most. In order to understand the relationship between the different traits and their influence on the presence of a circRNA, I developed several models using linear regression. Linear regression can be used to describe the association between one measurement (the response variable) and a set of other measurements (predictors).

For all annotated coding genes in each species, I collected data on their GC content, length (first to last exon), splicing potential (number of transcripts and exons), phastCons scores, expression levels and tissue-specificity. Furthermore, I counted the number of repeats in sense and antisense orientation (normalized by gene length) and used the previously calculated values for self-complementarity in sense and antisense orientation (**Chapter 2.6.1**) (**Table 6**). I then fitted a generalized linear model (GLM) to the data to describe the probability (p) of a coding gene to produce a circRNA (response variable). The previously listed structural and functional traits served as predictors for the model:

$$p(parental\ Gene) \sim GC\ content + gene\ length + ... + phastCons\ scores$$
$$with: family = logit(binomial)$$

Some of the predictors such as the GC content and gene length correlate with each other (known as multicollinearity), which can lead to a biased estimation of the effect size of individual predictors. To avoid this problem, I calculated the variance inflation factor (VIF) for each predictor. If the VIF indicated strong multicollinearity, only the predictor with the strongest effect was used for the final model. A detailed description of how this and the subsequent GLMs were fitted is provided in **Chapter 4.8**.

Consistently across all species, coding genes, which are located in genomic regions of low GC content, that span large genomic distances, which have a higher number of annotated exons and transcripts, a

Table 6: Structural and functional predictors for the GLM | The table provides an overview of the different structural and functional predictors that were used to construct the linear model. Data was taken from the indicated sources.

Predictor	Type	Source
Genomic length (bp)	Structural	Calculated
Number of exons	Structural	Ensembl
Number of transcripts	Structural	Ensembl
GC content	Structural	BioMart
Number of repeats in sense orientation to gene	Structural	Calculated
Number of repeats in antisense orientation to gene	Structural	Calculated
Self-complementarity in sense (%)	Structural	Calculated
Self-complementarity in antisense (%)	Structural	Calculated
PhastCons scores	Functional	Calculated
Expression levels (median FPKM across five tissues)	Functional	Brawand *et al.*, 2011 [22]
Tissue specificity (TSI)	Functional	Brawand *et al.*, 2011 [22]

high number of repeats in sense and antisense orientation and higher phastCons scores, are the most likely genes to produce circRNAs (**Table 7, Supplementary Table 7, Figure 22**). Interestingly, the influence of expression levels and tissue specificity of the parental gene is only mildly or not significant. In the following, I will discuss the significant predictors in more detail.

GC content and genomic length

In each species, low GC content is the strongest driver for the presence of a circRNA. The log odds ratios of scaled values range from -1.9 (opossum) to -0.8 in (rat). This is accompanied by an increased genomic length (log odds ratios from 0.3-0.5).

PhastCons scores

For mouse, rat and human, a high PhastCons score is the second strongest predictor for the presence

Table 7: GLM summary for parental genes | A generalized linear model was fitted to predict the probability of coding genes to be a parental gene ($n_{opossum}$=18,807, n_{mouse}=22,015, n_{rat}=11,654, n_{rhesus}=21,891, n_{human}=21,744). Predictors are sorted by their log odds ratio and only shown, if they were significant in at least two species or within primates/rodents. A summary of all log odds ratios and their p-values is provided in Supplementary Table 7. *Abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

Response	Predictor	Log odds range	Significance	Species
Parental gene	GC content	-1.9 to -0.8	***	md|mm|rn|rm|hs
Parental gene	PhastCons scores	0.6-0.9	***	mm|rn|hs
Parental gene	Genomic length	0.3-0.5	***	md|mm|rn|rm|hs
Parental gene	Self-complementarity in antisense	0.2-0.6	***	md|mm|rn|rm|hs
Parental gene	Exon and/or transcript count	0.2-0.4	***	md|mm|rn|rm|hs
Parental gene	Self-complementarity in sense	0.1-0.3	***	md|mm|rn|hs
Parental gene	Expression	0.1	**	md|rn

of a parental gene (log odds ratios from 0.6-0.9). PhastCons scores belong to the functional group of predictors indicating that parental genes are under evolutionary constrains.

Repetitive structures

I approximated the number of repetitive structures in two different ways. First, I calculated for each gene the percentage of nucleotides that were inloved in self-complementarity. Second, I counted the number of repeats in a gene normalized by its genomic length (first to last exon). The repeat annotation might be biased by the quality of the genome, and therefore, approach one can be more precise when approach two does not work. In all five species, the number of repeats and self-complementarity correlate positively with the presence of a circRNA. The correlation for self-complementarity in antisense orientation (0.2-0.6) is higher than for self-complementarity in sense orientation (0.1-0.3).

Transcript and exon count

In all species, the number of annotated exons and transcripts is positively correlated with the presence of a parental gene. Values for the log odds ratio range from 0.2-0.4.

A small number of circRNAs has been associated with a specific function. In human, 16 of them are found in coding genes and according to the GLM, are associated with high probabilities to produce circRNAs (**Supplementary Figure 2**). Coding genes with a high probability score in the GLM, may thus not only be likely to produce circRNAs, but those circRNAs could also be prioritized in functional studies.

2.6.3 GLMs for parental hotspots and shared circRNA loci

Next, I used the same approach to fit a model for parental genes to predict which parental genes are circRNA hotspots. For this model, I included also information on the presence of the circRNA locus in other species. The presence in multiple species is based on the previously defined level-2 classification of circRNA overlap in multiple species (**Chapter 2.5.1**). A therian circRNA locus for example, occurs in opossum and three other species, while a species-specific hotspot occurs in only one species. Interestingly, the overlap between species is the strongest predictor for the presence of a hotspot (log odds ratios from -2.0 to -0.8 for species-specificity). The correlation is negative,

indicating that genes, which are parental in only one species, have a lower probability of being a hotspot gene. In contrast, parental genes that occur in many species are more likely to be a hotspot (**Table 8, Figure 21A)**. The GC content of the parental gene influences the presence of a hotspot. However, the effect is milder than in the previous model that was fitted for parental genes (log odds ratios from -0.6 to -0.2) (**Table 8, Figure 21B)**. The presence of repeats seems to be less important. Self-complementarity in sense orientation is only significant in three species, and self-complementarity in antisense is only significant in mouse. A summary for all log odds ratios is provided in **Supplementary Table 8**. In this analysis the presence of a hotspot was predicted based on features from the parental gene. In **Chapter 2.8**, I will repeat a similar analysis, but instead on properties of the circRNA locus itself.

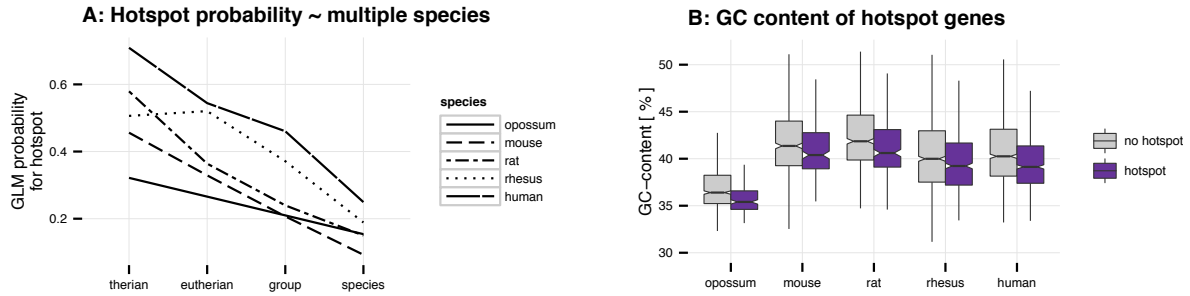


Figure 21: GLM for hotspots | A: Hotspot probability ~ multiple species. The probability of parental genes to be a hotspot is plotted against the hotspots' overlap with 1 (species-specific), 2 (group), 3/4 (eutherian) or 5 (therian) species. **B:** GC content of parental hotspot genes. The GC content of the parental gene in different species is plotted depending on whether it is a hotspot (purple) or not (grey). *Outliers for panel B were removed for plotting*

Last, I analyzed parental genes for components that could influence the presence of a circRNA locus in multiple species. The only shared predictors between all species are low GC content and genomic length, although the effect is milder than for other responses (log odds ratios from -0.6 to -0.3 and from 0.2-0.3 respectively). Species-specific circRNA loci in contrast, are characterized by higher GC content (log odds ratios from 0.2-0.6), lower phastCons scores, fewer transcripts and shorter genomic length (**Table 8, Supplementary Table 9**). None of the predictors for species specificity is present in all five species, and in general, correlations are opposite to the previous trends detected. A more refined analysis, which is trying to answer the question of circRNA presence in multiple species, will be discussed in **Chapter 2.8**.

Table 8: GLM summary for parental hotspot genes | A generalized linear model was fitted to predict the probability of parental genes to be a hotspot ($n_{opossum}$=884, n_{mouse}=858, n_{rat}=983, n_{rhesus}=1704, n_{human}=2058). For each response, predictors are sorted by their log odds ratio and only shown, if they were significant in at least two species or within primates/rodents. A summary of all log odds ratios and their p-values is provided in Supplementary Table 8 and 9. *Abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

Response	Predictor	Log odds range	Significance	Species
Hotspot gene	Number of species	-2 to -0.8	***	md|mm|rn|rm|hs
Hotspot gene	GC content	-0.6 to -0.2	***	md|rn|rm|hs
Shared circRNA locus	GC content	-0.6 to -0.3	***	md|mm|rn|rm|hs
Shared circRNA locus	Genomic length	0.2-0.3	**	rn|rm
Species-specific circRNA locus	GC content	0.2-0.6	***	md|rn|rm|hs
Species-specific circRNA locus	PhastCons scores	-0.4 to -0.3	***	mm|rn|hs
Species-specific circRNA locus	Transcript count	-0.3 to -0.2	**	mm|rm|hs
Species-specific circRNA locus	Genomic length	-0.4 to -0.1	**	md|rn|rm

2.6.4 Summary for all linear models

The use of linear models to incorporate the different structural and functional measurements associated with circRNA parental genes has provided strong evidence that the presence of a circRNA is correlated with several structural properties of the parental gene. Parental genes are GC-low, large, have high splicing potential and are enriched in repeats. Interestingly, a high phastCons score is a strong positive predictor for the presence of parental gene suggesting that parental genes could be under evolutionary constrains. The previously listed structural components determine also which genes among all parental genes are circRNA hotspots. More importantly, the presence of a hotspot is strongly associated with the number of species in which a gene produces a circRNA. If the gene is a parental gene in many species, it is more likely to be a hotspot gene. Last, the linear models suggest that parental genes of shared circRNA loci are not shared due to higher expression levels or high conservation levels of the parental gene, but mainly due to low GC content. Species-specific circRNA loci in contrast are found in parental genes with opposite features. They possess higher GC content, are short, give rise to fewer transcripts and are associated with lower phastCons scores suggesting that they are located in the lower tail of the probability distribution for parental genes (**Figure 22**). The here-presented results support the hypothesis that certain genomic properties facilitate the biogenesis of circRNAs. Orthologous genes are more likely to possess a similar structure, which could lead to the presence of similar circRNAs across species.

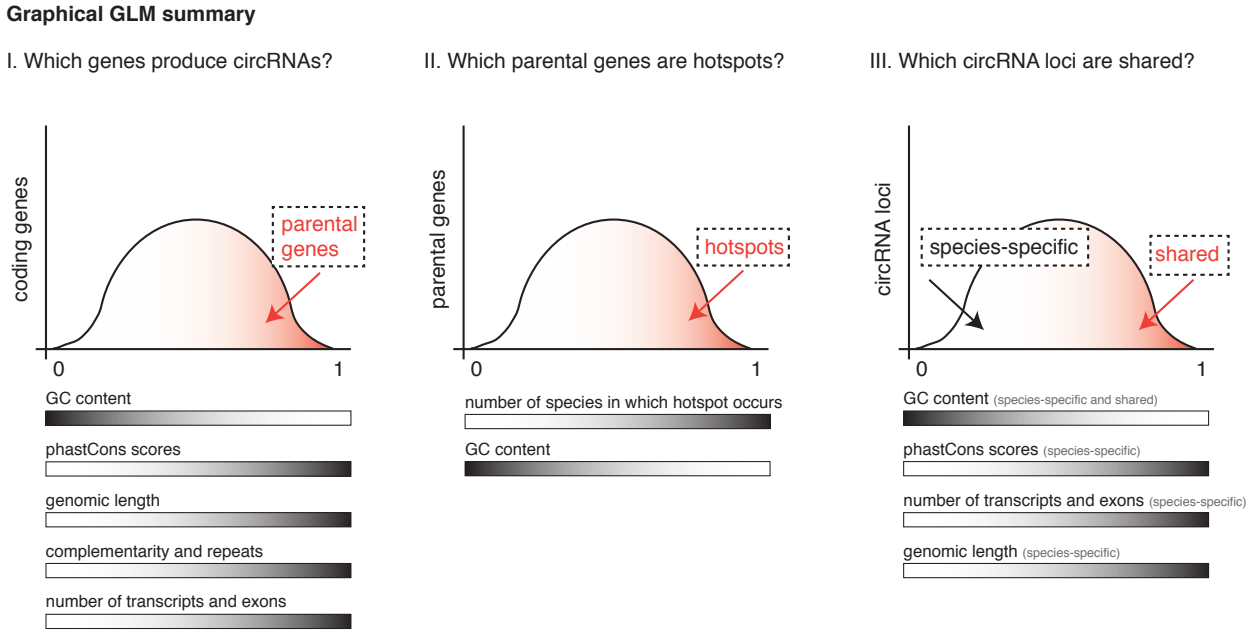


Figure 22: Graphical overview of the different GLMs | The probability distribution for parental genes (I), hotspot genes (II) and shared/species-specific circRNA loci (III) is outlined. An increase in color intensity reflects higher probabilities for the occurrence of the event (white -> red) or higher values of the different predictors used to fit the linear model (white -> black). Errors indicate where in a distribution the different events are found. All predictors (GC content etc.) are sorted according to importance for the event. In summary, circRNA parental genes are GC-low, have high splicing potential and are long. They are enriched in repeats and have high phastCons scores (I). Parental hotspot genes have low GC content and are present in multiple species (II). Shared circRNA loci are found in GC-low parental genes. Species-specific circRNA loci are produced by short parental genes, with fewer exons and lower phastCons scores (III).

2.7 CircRNAs and repeats

2.7.1 Enriched repeat families in flanking introns and RVCs

As shown in the previous chapter, the likelihood of a coding gene to produce a circRNA depends on its GC content and the surrounding repeat frequency. Furthermore, the presence of a circRNA in multiple species is explained neither by high phastCons scores nor high expression levels of the parental gene, but instead by structural commonalities between the parental genes. At this point, it remains unclear whether shared circRNAs emerged independently from each other, or whether there was an ancestral circRNA in all species. Driven by changes in the repeat environment, latter may have changed the backsplice sites throughout time (turnover), which could explain the low number of circRNAs sharing the same first and last exon (**Figure 16**). To further investigate the two scenarios of common or independent circRNA origin, I decided to analyze the repeat environment of parental genes in more detail. In case of an ancestral circRNA, the flanking introns of shared circRNAs should contain transposable elements common between species. In addition, one could expect that the dominant circRNA in each hotspot represents the most ancestral circRNA. Because the circRNA has been present over a long period of time, expression levels stabilized since its emergence leading to the observed dominance. In case of independent emergence, the presence of species-specific integration events between orthologous parental genes would independently create a repetitive genomic environment and support the hypothesis of a parallel circRNA origin. Dominance is now explained by recent integration, because repeat pairing is strong due to a lack of degradation and insufficient defense mechanisms of the host genome.

To be able to connect the observed repeat elements directly with the circRNA and to distinguish the two hypotheses (common or independent circRNA origin) from each other, I focused on the flanking introns only. I generated a control dataset of approximately the same number of introns that was matched for intron length. For the interpretation of results, I assumed that in order to influence the presence of a circRNA, repeats should fulfill the following criteria: They need to be upstream and downstream of the circRNA (1) and should come from the same repeat family (2) to form stable dimers (3) by reverse-complement base pairing (4) (**Figure 23**).

In comparison to the background dataset, flanking introns are enriched in repeats in sense- and antisense-orientation to the gene (sense: 2.0-2.9x, antisense: 2.1-3.2x) (**Figure 24A**). If the background data set is additionally corrected for GC content, the enrichment of repeats in the flanking introns is less pronounced, but still significant (data not shown).

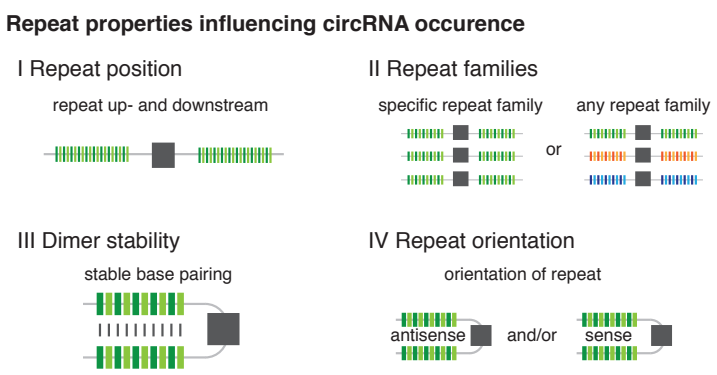


Figure 23: Repeat properties influencing circRNA occurrence | To influence the presence of a circRNA, repeats should be upstream and downstream of the circRNA (I), from the same repeat family (II) and form stable dimers (III) by reverse-complement base pairing (IV). It is unknown to what extent condition II and IV are true. In this scheme, grey boxes represent the circRNA body, while striped boxes indicate a repetitive element.

Several studies have shown that circRNA flanking introns are enriched in repeats belonging to the SINE family [64, 65, 67]. Therefore, I decided to investigate the abundance and enrichment of SINE elements in the flanking introns of all species. Rodents have five major SINE families: B1 (in RepeatMasker annotated as Alu), B2, B4, ID and MIR. In primates, there are two major SINE families: Alu and MIR. Based on their age, I divided the Alu family into three subgroups (AluJ, AluS, AluY). For opossum, I only distinguished between MIR-related and MIR-unrelated SINE elements. In all five species, MIR elements can be abundant, but are only lowly enriched. There are no strong differences in enrichment between the remaining TE families (**Figure 24B' and C', Supplementary Figure 3A', B' and C'**). To understand if specific repeats drive the formation of circRNAs, I compared the abundance and the enrichment of individual members of each SINE family in the flanking and background introns. In all species, the abundance of TEs correlates with their reported activity. In opossum, SINE1_Mdo - a recent and (still) active TE is the most abundant and enriched (**Supplementary Figure 3A**). In mouse, the most abundant and enriched TEs are B3, B4, B1_Mus1 and B1_Mus2 (**Figure 24B"**). In rat, B3 and B4 are similarly abundant, however B1 family members do not occur often. Instead, recently active ID family members are more frequently found and enriched (**Supplementary Figure 3B"**). In primates, the most frequent and enriched TEs in the flanking introns are AluJ and AluS. Both elements are characterized by high amplification rates in respect to AluY (AluJ: 20,000 copies/myr, AluS: 2000 copies/myr, AluY: 100-500 copies/myr) (**Figure 24C", Supplementary Figure 3C**) [116].

To assess which repeats are forming dimers, I annotated the detected reverse-complementary and complementary alignments identified by megaBLAST with RepeatMasker. To be sure of the repeat presence, repeats had to overlap with at least 50% of their length with the megaBlast alignment. In opossum, the most prevalent repeat pair is the SINE1_Mdo dimer. In mouse, several dimers within

the B1 family (B1_Mus1, B1_Mus2, B1_Mm) and the B2 family (B2_Mm2) are found. Rat is characterized by dimers of the ID family (ID_Rn1, ID_Rn2). In rhesus macaque and human, pairs between Alu elements are frequently found. While in rhesus the prevalent pairs consist of AluSx and others, in human the most common pairs are between AluSx1 and AluSx (**Figure 25A**). A similar trend is also present in sense orientation (**Supplementary Figure 4**). Interestingly, all enriched repeats are not only species-specific or lineage-specific, but have undergone recent amplification rounds and are thus young. Young repeats had less time to degrade and may thus be those that can base pair and loop the most efficiently. Because the DNA sequence varies less from the consensus, they are easier to annotate, which could partially drive the observation. In addition, several alignments did not overlap with annotated repeats under the chosen thresholds (in antisense 13.9% in opossum, 7.5% in mouse, 7.0% in rat, 3.0% in rhesus and 3.3% in human). They might either be strongly degraded and therefore, they are not annotated, overlap with less than 50% with the alignment, or present simple DNA repeats.

Next, I analyzed the pairing capability within sense and antisense dimers of all TEs to confirm that recent TEs are more likely to pair with each other than older TEs. For each TE, I received the consensus sequence from Repbase and used RNAcofold from the ViennaRNA package to fold TEs on each other [117, 118]. RNAcofold calculates the minimal free energy (MFE, unit in kcal/mol) for the secondary structure of RNA and DNA dimers. The MFE can be used as characteristic value to describe the dimer stability - the lower the value the more stable. In all species, the secondary structure of antisense dimers is more stable than in sense dimers (**Figure 25B**). Furthermore, the most stable structures are found in dimers composed of the same TE. However, TEs can also form stable structures with other TEs from the same family or with TEs from another family (**SupplementaryTable 10**). This is of particular interest for mouse, rhesus and human, because for these species some of the TE families are very large (B1 family: 10 TE members, Alu family in human: $>$ 15 TE members). Therefore, dimer formation, which is important to form the backloop during circRNA biogenesis, can occur across a broad spectrum of TEs.

The binding landscape is complex and different TE dimers can be created between different TEs belonging to different families. In general, antisense-binding is preferred, but sense-binding might also be of importance. With RNAcofold, I approximated dimer stability *in-silico* using consensus sequences. However, the high abundance of TEs in close proximity to each in the circRNA flanking introns might favor different dimers. In addition, the local DNA structure and environment (e.g. modifications, chromatin structure) may influence the binding probability and strength and

could create spatio-temporal binding preferences. In summary, analyses on TE dimer frequency and stability suggest that young TEs are influencing circRNA formation and therefore, new integration events may influence which circRNA will be produced in a hotspot. In the next chapter, I will therefore analyze and compare the repeat landscape of circRNAs depending on their overlap with other species and their dominance in a hotspot.

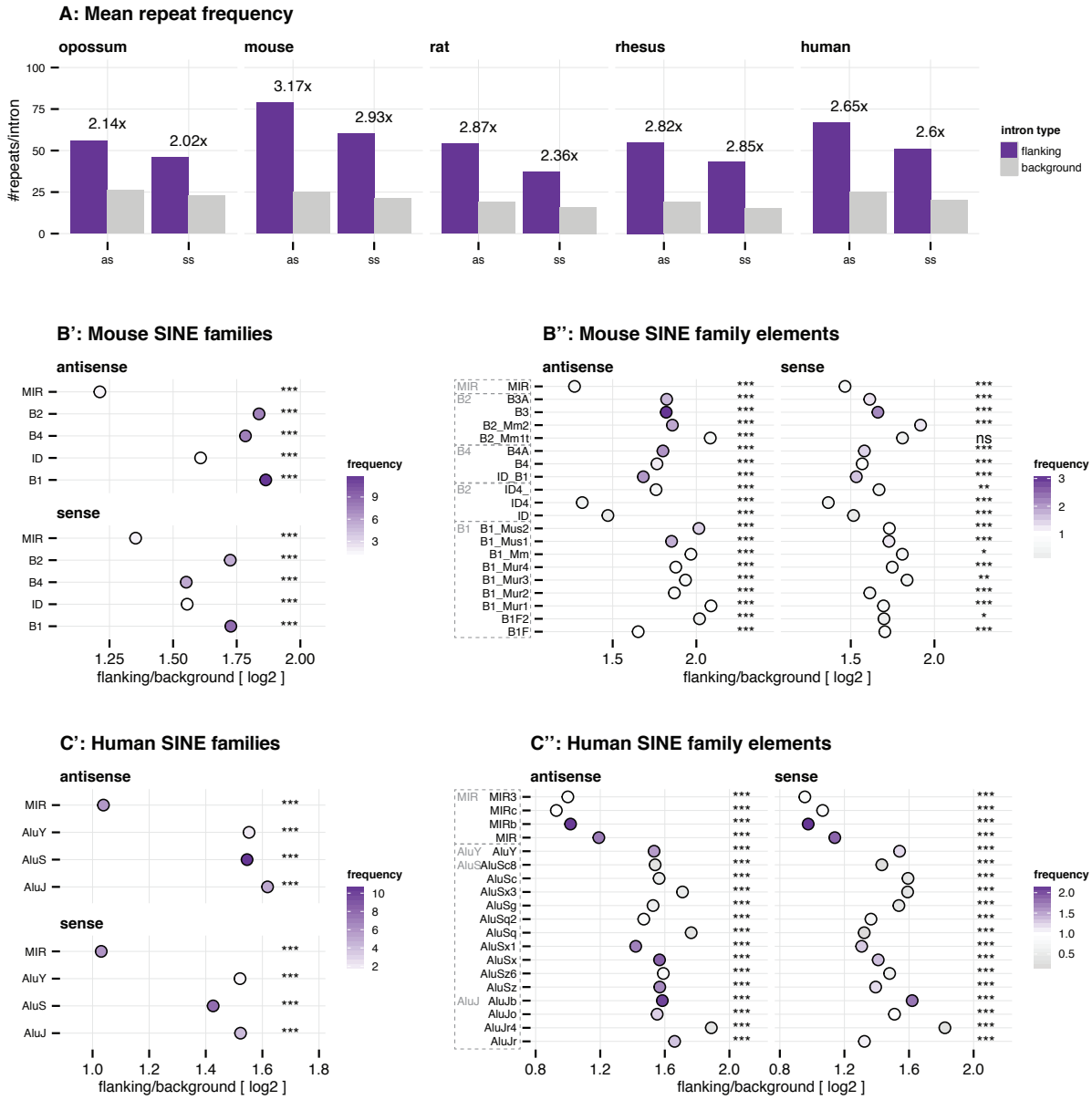


Figure 24: Repeat frequency in flanking and background introns | A: Mean repeat frequency. Figure shows the total number of repeats in flanking (purple) and background introns (grey). The enrichment of repeats in flanking introns was calculated for each species in sense and antisense orientation. **B':** Mouse SINE families. Plotted is the log2-enrichment for different repeat families in antisense and sense orientation. Increase in color intensity reflects the mean number of repeats detected in all flanking introns. Significance was estimated with a one-tailed Mann-Whitney U test. **B'':** Mouse SINE family elements. Plotted is the log2-enrichment for individual SINE family members in antisense and sense orientation. Increase in color intensity reflects the mean number of repeats detected in all flanking introns. Significance was estimated with a one-tailed Mann Whitney U test. Family members that were not significant in sense and antisense orientation were removed from plot. **C':** Human SINE families. Plot was generated as described in B'. **C'':** Human SINE family elements. Plot was generated as described in B''. *Abbreviations: as = antisense, ss = sense. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

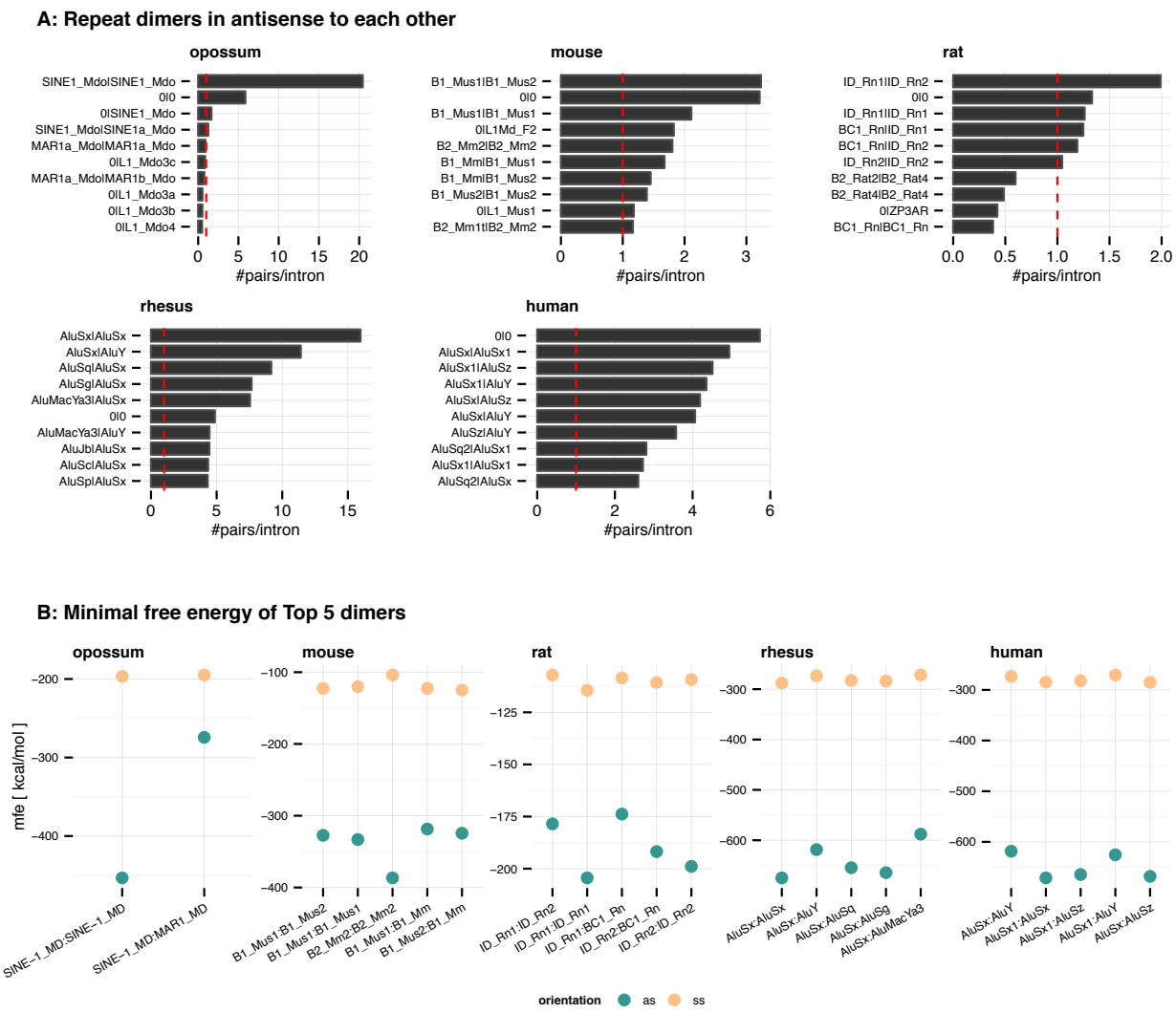


Figure 25: Reverse-complement repeat enrichment | A: Repeat dimers in antisense to each other. Frequencies of the Top 10 repeat dimers found in flanking introns of each species. Red line was set at position 1 to indicate which repeats occur less or more often than once in the flanking introns. "0" in a dimer name corresponds to alignments for which no overlapping repeat was found after the intersection with the RepeatMasker annotation. **B:** Minimal free energy of Top 5 dimers. MFE (kcal/mol) of the secondary structure of the Top 5 dimers was calculated with the RNAcofold function from the ViennaRNA package. MFE was assessed for sense-pairing (yellow) and antisense-pairing. As approximation of the TE sequence, the consensus sequence from Repbase was used. "0|0" dimers were not assessed, because no consensus sequence exists. *Abbreviations: as = antisense, ss = sense.*

2.8 Hotspots

2.8.1 Linear regression models for hotspot presence and depth

Given the high abundance and enrichment of TEs in the flanking introns of circRNAs, I asked whether TEs could also explain how many circRNAs are found in a hotspot (depth) and which circRNA will be dominant (most highly expressed) in a hotspot. Similar to the previous analyses in **Chapter 2.6.2**, I again used linear regression to adress these questions. In addition to the repeat frequencies, I incorporated information on the splicing strength of each circRNA. To approximate the splicing strength, I calculated the intron and exons GC content (last intronic 250 nt, first exonic 50 nt), the amplitude and ΔG at the backsplice site. ΔG was calculated for the first 100, 250 and 500 nt at each splice site. I also incorporated information on the hotspot and circRNA presence in other species as calculated in **Chapter 2.5.1**. All predictors are listed in **Table 9**.

In **Chapter 2.6.2**, I estimated the presence of a hotspot within parental genes from structural components describing the parental gene. To be more precise, I focussed now on properties that are more specific to the circRNA locus itself such as the length of the locus or the quality of the splicing signals. The linear model was run as follows:

$$p(presence\ of\ hotspot) \sim GC\ content\ +\ length\ +\ ...\ + \Delta G$$

$$with: family = logit(binomial)$$

Table 9: GLM predictors for hotspots presence, depth and circRNA dominance | All predictors were calculated using BioMart or the sequence information itself. They could influence either looping of the mRNA or the splicing of the circRNA.

Predictor	Potential influence on:	Response
GC content	Looping	Hotspot presence, depth
Genomic length of hotspot	Looping, splicing	Hotspot presence, depth
Total number of repeats in flanking introns	Looping	Hotspot presence, depth
Median ΔG of splice sites	Splicing	Hotspot presence, depth
Intron GC content	Splicing	Dominance
Exon GC content	Splicing	Dominance
Acceptor amplitude	Splicing	Dominance
Donor amplitude	Splicing	Dominance
Length of transcript	Splicing	Dominance
Number of exons	Splicing	Dominance
Number of repeats in flanking introns	Splicing	Dominance
Mean distance of repeat pair to splice site	Splicing	Dominance
ΔG of splice site (100, 250, 500 nt)	Splicing	Dominance

Table 10: GLM summary for hotspot presence and depth | A generalized linear model was fitted to predict the probability of hotspot presence between all circRNA loci ($n_{opossum}$=1049, n_{mouse}=1024, n_{rat}=1285, n_{rhesus}=2169, n_{human}=2759) and the depth of a hotspot ($n_{opossum}$=203, n_{mouse}=234, n_{rat}=305, n_{rhesus}=605, n_{human}=846). For each response, predictors are sorted by their log odds ratio and only shown, if they were significant in at least two species. A summary of all log odds ratios and p-values is provided in Supplementary Table 11. *Species abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

Response	Predictor	Log odds range	Significance	Species
Hotspot	Number of repeats	1.9-3.5	***	md|mm|rn|rm|hs
Hotspot	Genomic length	1.0-1.3	***	md|mm|rn|rm|hs
Hotspot	Multiple species	0.7-1.3	***	mm|rn|rm|hs
Hotspot	GC content	-1.1 to -0.2	***	md|rn|rm|hs
Hotspot	ΔG	0.1-0.2	**	md|rn|rm
Hotspot depth	Number of repeats	0.1-0.2	***	md|mm|rn|rm|hs
Hotspot depth	Hotspot length	0.1-0.2	***	md|mm|rn|rm|hs

In agreement with the GLM for hotspots in **Chapter 2.6.2**, the presence of a hotspot among all parental genes correlates with low GC content (log odds ratios from -1.1 to -0.2) and its presence in multiple species (log odds ratios from 0.7-1.3). However, the strongest positive predictor is now the frequency of TEs (log-odds ratios from 1.9-3.5). In addition, the genomic length of a circRNA locus correlates positively with the hotspot presence: If circRNAs cover a large part of the parental gene, this area is more likely to produce several circRNAs (log odds ratios from 1.0-1.3). Furthermore, in opossum, mouse and rhesus, a high median ΔG at splice sites correlates positvely, although only mildly, with the presence of a hotspot (log odds ratios from 0.1-0.2).

The depth of a hotspot was analyzed for all hotspots that produce between two and five circRNAs. Hotspots with a higher number of circRNAs were not incorporated to avoid biases. The following model was used:

$$p(number\ of\ circRNAs/hotspot) \sim GC\ content\ +\ amplitude\ +\ ...\ + \Delta G$$
$$with: family = logit(poisson)$$

There are no strong predictors associated with the depth of a hotspot. The depth is mildly influenced by a higher number of repeats (logg odds range from 0.1-0.2) and increased hotspots length (logg odds range from 0.1-0.2) (**Table 10, Supplementary Table 11**). However, the number of available loci to analyze is low due to the restriction for the number of circRNAs in a hotspot (between two and five), which could limit the power of this analysis.

2.8.2 Properties of the dominant circRNA in each hotspot

The previous models indicated that the presence of a hotspot is strongly influenced by the number of repeats. Therefore, I asked whether the number of individual repeats in the flanking introns correlates with the dominance of a circRNA in the hotspot. To assess dominance, I sorted circRNAs according to their expression levels in each hotspot and analyzed their rank (most highly expressed circRNA = rank 1, second most highly = rank 2, etc.). The reciprocal rank (1/rank) was used to create a positive correlation between rank and expression levels. In contrast to the previous analyses, I decided to use a linear mixed model (LMM) for this analysis. Different hotspots have different genomic backgrounds and therefore, contribute different noise levels for which the LMM can account:

$$p(1/(rank\ of\ circRNA)) \sim repeats\ +\ genomic\ length\ +\ ...\ +\ age$$
$$with:\ background\ effect = hotspot$$

The linear mixed model suggests only a weak link between circRNA dominance and two components: Presence of the circRNA in other species and genomic length. Dominant circRNAs are more often found in multiple species and span a smaller genomic distance. However, the effect is small and interpretation should be done carefully (log odds ratio between -0.1 to -0.02 for both predictors). Interestingly, the LMM does not indicate a correlation between the strength of a splicing signal (e.g. GC amplitude or ΔG) and the rank of a circRNA (**Table 11, Supplementary Table 12**).

Given the low effect size of circRNA presence in multiple species on circRNA dominance, I decided to confirm this result independently. I compared the expected and observed overlap between species of the dominant circRNA and other circRNAs from the same hotspot (overlap ratio). The dominant circRNA can be present in more species (group 1), the same number of species (group

Table 11: LLM summary for circRNA dominance | A linear mixed model was fitted to predict circRNA dominance in a hotspot ($n_{opossum}$=203, n_{mouse}=234, n_{rat}=305, n_{rhesus}=605, n_{human}=846). In each hotspot, circRNAs were ranked by the expression levels (rank 1 = highest expression, rank 2 = second highest expression etc.). Analysis was restricted to hotspots with two to five circRNAs. For each response, predictors are sorted by their log odds ratio and only shown, if they were significant in at least two species. A summary of all log odds ratios and p-values is provided in Supplementary Table 12. *Species abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

Response	Predictor	Log odds range	Significance	Species
Dominance	circRNA age	0.13	***	mm|rn|rm|hs
Dominance	circRNA genomic length	-0.02 to -0.04	***	md|mm|rm|hs

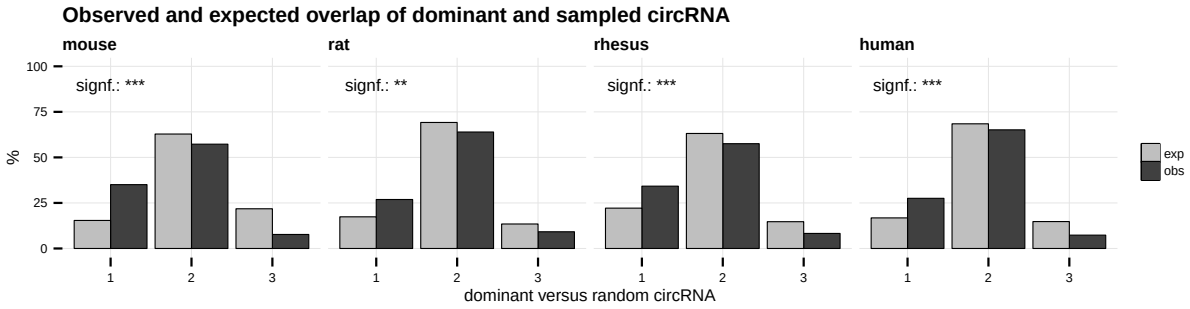


Figure 26: Observed and expected overlap of dominant and sampled circRNA | Expected overlap ratio (light grey) was estimated by sampling two circRNAs from each hotspot and comparing their overlap with each other. Observed values (dark grey) were calculated by comparing the dominant circRNA with a second circRNA sampled from the same hotspot. The ratio can be divided into three groups: In group 1 the dominant circRNA occurs in more species than the sampled circRNA, in group 2 they occur in the same number of species and in group three, the dominant circRNA occurs in fewer species. The difference between expected and observed overlap ratios was estimated with a one-tailed and paired Mann-Whitney U test, all p-values were smaller than 0.001. *Abbreviations: exp = expected, obs = observed*

2) or in fewer species (group 3) when compared to a second circRNA from the same hotspot. To estimate the expected overlap ratio, I sampled two circRNAs from each hotspot and compared the number of species in which the first circRNA occurred with the number of species, in which the second circRNA occurred. The observed ratio was calculated using the dominant circRNA and a sampled circRNA from the same hotspot. The analysis was restricted to primates and rodents, because based on the overlap definition of **Chapter 2.5.1**, opossum circRNAs can only be shared in all species or be opossum-specific, which might create a bias in the calculations. For all rodents and primates, the overlap ratio is significantly higher than expected, suggesting that the dominant isoform is more often shared across multiple species (one-tailed and paired Mann-Whitney U test, $p < 0.001$) (**Figure 26**).

Dominant circRNAs overlap in more species than non-dominant circRNAs from the same hotspot. In addition, the depth of a hotspot is influenced by a high number of TEs, and young TEs from the SINE family are enriched in reverse-complementary regions in the flanking introns (**Chapter 2.7**). To understand if there is a correlation between young repeats and the dominant circRNA, I decided to analyze the repeat environment (repeat abundance, repeat family and age) for dominant and non-dominant circRNAs in more detail. For each circRNA, I counted the number of individual SINE family members. I then compared their abundance in the flanking introns of the dominant isoform and a randomly sampled circRNA from the same hotpot.

The results of this analysis are less clear than for previous analyses. In mouse for instance,

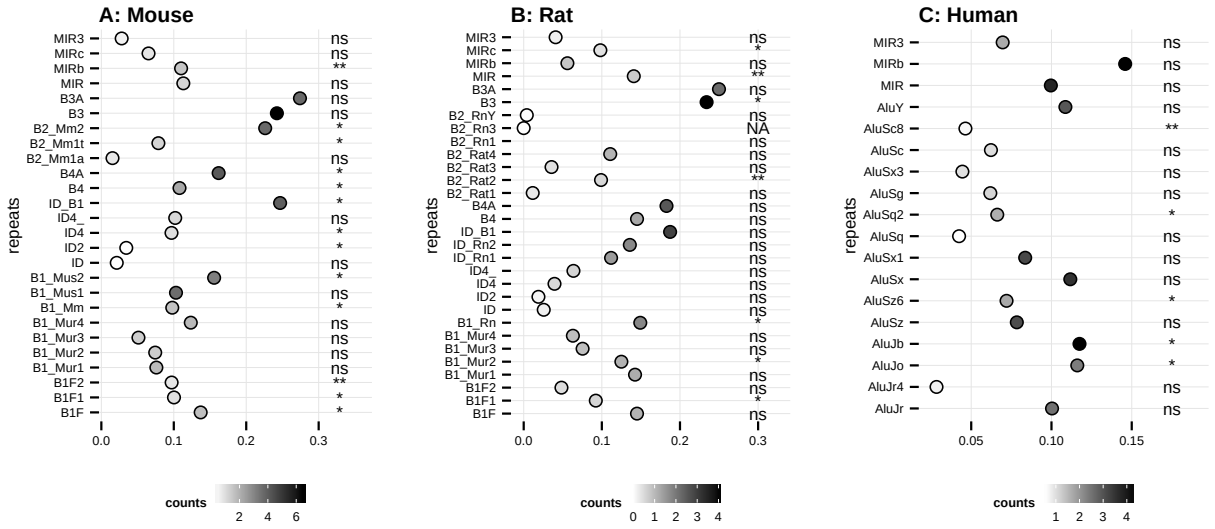


Figure 27: TE environment of dominant circRNAs | A: Mouse. The frequency of different TEs in flanking introns of dominant and randomly sampled circRNAs from the same hotspot was assessed. Sampling was repeated 1000 times for each dominant circRNA. The mean species overlap of the sampled circRNA was used to calculate the enrichment. Plotted is the log2-enrichment for different repeat families. An increase in color intensity reflects the mean number of TEs detected in the flanking introns of dominant circRNAs. Significance was estimated with a one-tailed and paired Fisher's t test. **B:** Rat. **C:** Human. *Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

B2_Mm2, B2_Mm1t, B1_Mm and B1_Mus2 are enriched (**Figure 27A**). The enrichment is small, but significant. However, in rat the ID family members do not occur more frequently (**Figure 27B**). In human, members of the AluSx family are also not enriched. In contrast AluJo and AluJb are found more often (**Figure 27C**). In opossum and rhesus macaque, none of the TEs are significantly enriched (**Supplementary Figure 5**). In summary, there might be a tendency for younger repeats to be enriched in the flanking introns, but it is weak. In addition, the use of linear regression may not be the right choice due to a low sample size.

2.9 A biological model for the production of circRNAs

The leading assumption in the circRNA field is that circRNAs are a conserved, putative functionally important class of RNAs. The presence of circRNAs in multiple species is one of the most frequently arguments to support their importance. However, the underlying assumptions for this argument were not sufficiently tested. Based on the here-presented analyses, following observations were made:

1. CircRNA parental genes present a specific subgroup of coding genes: They are GC poor, long and have a high potential for alternative splicing. In addition, they are highly enriched in repeats.

2. Repeats, which are enriched in parental genes, belong to species-specific SINE families.

3. The formation of the loop required for circRNA biogenesis is supported by recently active repeats. Because they were recently active, these repeats might be the least degraded and form the strongest loops.

4. In circRNA hotspots, the dominant circRNA is in average more often shared between species than randomly sampled non-dominant circRNAs from the same hotspot.

5. In addition, dominant circRNAs might be flanked by more and younger repeats than non-dominant repeats.

Based on these observations, I propose the following model for circRNA biogenesis and overlap between species: Coding genes that are GC-low and large are predisposed for the integration of TEs in their introns. Because of the increased intron size, integrated TEs are neutral or only mildly deleterious for the gene. The presence of a TE can lead to the formation of an integration hotspot as described for intergenic regions by Levy *et al.* [28]. The combination of low-GC, long introns, higher splicing capability and many TEs creates a genomic context in which the production of circRNAs is facilitated. The presence of multiple repeats supports looping and can lead to the formation of circRNAs via backsplicing. CircRNAs for which the loop formation is strong will be dominant in a hotspot. Strong loops are formed by young repeats, which are the least degraded.

Furthermore, shared circRNAs are found in orthologous parental genes that possess similar structure and thus provide independently from each other a similar genomic context for circRNA production. Because circRNAs depend on recently integrated and species-specific TEs, most shared circRNAs emerged in parallel and not from a common ancestral circRNA.

Thus, most circRNAs are not conserved. They present a shared side-product, which is driven by the recent integration of species-specific TEs into structurally predisposed orthologous coding genes.

3 Discussion

3.1 Improvements and drawbacks in the circRNA detection pipeline

CircRNAs are characterized by their low expression levels. It requires several filtering steps and careful analysis to distinguish them from the technical (and biological) noise generated by RNA-seq protocols. In a standard read mapping protocol, about 10% of reads are unmapped due to differences between the sample and the reference genome, repetitive regions in the genome, sequencing errors or non-canonical splicing behavior (as observed for circRNAs) [119]. Hence, reads that are indicative for circRNAs, need to be recycled from the unmapped fraction of reads. Normally, this is done by the remapping of anchors generated from the terminal 3'- and 5'-end of a read. If the left and right anchor map in a non-linear order, are within a certain distance to each other and are extendable to the full read length, then the read may come from a backsplice site [120]. In a very naive approach, every read that fulfills these criteria could be a backsplice candidate. Many studies have reported very high numbers of circRNAs, without adequate filtering and validations steps and it is likely that these datasets contain high numbers of false-positives. Consequently, many findings might be biased [92, 97]. It is important to be aware of the technical problems and variables that can influence the detection of BSJ reads. Therefore, I will briefly discuss these problems in the next paragraphs and describe, how I have tried to minimize them.

Many circRNA studies use cell lines, only two replicates and avoid the generation of RNase R treated samples due to higher sequencing costs [97]. CircRNAs are expressed at levels close to the detection limit. Hence, it is important to avoid experimental designs with high noise levels (e.g. cell lines, few replicates). In the here-presented dataset, I therefore used real tissues from individual animals of approximately the same age, worked with triplicates (three individuals per condition) and made use of RNase R treatment to distinguish circRNAs from noise. In addition, paired-end sequencing data was used to increase the information content around the splice site by the mate mapping information.

I have chosen an anchor length of 20 bp, which is a compromise between the mapping precision (that decreases, the smaller the read) and the identification of an event (data not shown). An increase in length will lead to an increase in the mapping quality, but on the expense of loosing candidates for which the read covers a junction with less than the x chosen base pairs. If the anchor maps equally well to several regions in the genome, many mappers will randomly choose one of the mapping positions and report only this one if not indicated differently. As a consequence, many true events are missed, because the mapper picked the incorrect position. Analyzing all possible

mappings requires more time and a more careful statistical analysis of the different mapping events. I did not incorporate this step at the time of the pipeline development, because the main focus of my thesis was not the development of an highly accurate identification pipeline, but the downstream analysis of predicted circRNA candidates.

I have set the distance between the mapped ends of an anchor pair to 100 kb, which reflects the average gene length. Increasing the distance to 500 kb can help to identify further candidates. After 500 kb, the number of detected backsplice sites remains relatively stable (data not shown). The genomic region upstream and downstream of the backsplice site is often repetitive, which increases the problem of finding the exact breakpoint of the read. To overcome this problem, several detection tools trim the BSJ read to the next canonical splice site. However, it has been acknowledged that circRNAs can also use non-canonical splicing motifs and therefore, I trimmed overlapping reads to known canonical and non-canonical splicing motifs [54, 57]. Reads that did not pass this filter were discarded. More than 95% of the final BSJs are flanked by canonical splice sites (data not shown).

I used an enrichment approach to discard putative BSJs for which the read coverage did not increase after RNase R treatment. In my pipeline, I used the same enrichment threshold for all coverage levels. However, Szabo *et al.* showed recently that the confidence interval for RNase R enrichment changes as a function of the read count. The confidence interval for the RNase R enrichment in samples with low read coverage is broader than for deeply sequenced samples. BSJs could thus look like they were depleted if one compares a lowly sequenced RNase R treated and a deeply sequenced untreated sample with each other [97]. RNase R treated libraries are more difficult to sequence and are biased due to the high number of multi-mapped reads. Therefore, their true sequencing depth is difficult to estimate. By choosing a fixed cut-off, I may have lost several candidates.

Finally, I implemented a second remapping step to increase the coverage on the backsplice site, which might partially overcome the enrichment bias. However, some BSJs may not have been discovered due to the initial 20 bp threshold for mapping.

I reconstructed potential circRNA isoforms by creating splicing graphs based on the enrichment and decrease of splice sites within and outside the BSJ after RNase R treatment. The reconstruction is imprecise for circRNA transcripts close and below 0.05 CPMs. For these expression levels, there are on average only 1-4 reads per BSJ (if at all) and it is impossible to deduce the transcript structure [97].

In summary, I have tried to be careful with the definition of a real BSJ, the quantification and transcript reconstruction, although several steps could be improved throughout the pipeline. Nevertheless, the pipeline is robust enough to detect high-confident circRNA candidates, which can be used for subsequent analyses.

3.2 Current controversies over circRNA function and conservation

Because circRNAs are found in orthologous genes across many species, it is widely assumed that they are a conserved and thus functionally important property of the genome. However, the presence of a feature in multiple species does not automatically imply functionality. The observed overlap could, for instance, simply be an overlap of transcriptional noise. Each gene has its own characteristic source and level of noise. Orthologous genes are often similar in their structure and expression profiles causing similar noise patterns. It is possible to speak of "orthologous noise" in this case, but claims on functionality should not be drawn without adequate biological experiments.

The dataset I presented in this study allows to address the question of circRNA functionality and conservation levels in greater detail. In the following, I would like to argue against some of the most frequent claims made in the circRNA research field by using the results of my own analyses.

3.2.1 Controversy 1 - CircRNAs are frequent and therefore important

The number of reported circRNAs in a given tissue ranges from about 100 to almost 40,000 (e.g. 103 circRNAs in human embryonic stem cells [121] versus 38,983 circRNAs in human frontal cortex [66]). The frequency varies dramatically between cell types, tissues and organisms without any obvious underlying pattern (**Table 12**). Based on these numbers, many studies have suggested that circRNAs are not only an abundant, but also important feature of the mammalian RNA landscape.

Based on my analyses, I hypothesize instead that the high number of genes producing circRNAs is explained by the properties parental genes exhibit. These properties are low GC content, long introns, a higher splicing capacity and a high number of repetitive elements, which is characteristic for genes in the isochores L1, L2 and H1, with L2-genes incorporating most of them. In human, approximately 53% of coding genes are found in these three isochores. L2 itself contains 25% of coding genes. The distribution is similar across the here-analyzed rodents and primates, while opossum coding genes occur mainly in the GC-low isochores L1 and L2 (**Table 13**).

Low GC content and long introns can allow for the integration, but also fixation of TEs. Levy *et al.* showed that TE integration hotspots exist in intergenic regions and that very often, TEs integrate

into already existing TEs [28, 42]. The fixation of a TE in a specific genomic locus suggests that first, the host genome is not strongly affected by the integration of the TE in this locus, and second that the TE has found a stable area, from which it can continue to amplify. Because of these advantages, other TEs are likely to be fixed in the same locus. The process of repetitive integration may finally lead to a TE hotspot. Genes with long introns may allow the formation of TE hotspots. Detrimental effects on transcription and splicing could be neglectable, if the integration event is located distantly enough from any regulatory signal. Zhang *et al.* indeed found that intronic repeats are under-represented in proximity to exon boundaries, but exhibit random integration patterns in distal areas [43]. L1 elements have a fixation bias for genomic regions with low GC content and in human, many Alu elements integrate into already existing L1 elements. Therefore, the integration bias of SINE elements into L1 elements, which are fixed in GC-poor regions, can cause an integration bias of SINEs into GC-poor regions [28]. The emergency of an intronic TE hotspot is thus a function of low GC content, large introns and the SINE-integration/LINE-fixation biases. The most likely gene candidates allowing for TE hotspots should thus be found in L1 and L2 isochores (**Figure 28A-B**).

TEs are important for the formation of circRNAs. L1 and L2 genes are enriched in circRNA parental genes, which is in line with the previous assumption of the TE integration bias into GC-low isochores. In human for instance, 20% of L1 and L2 coding genes produce a circRNA, while in contrast only 3.4% of coding genes in H2 are parental genes (**Table 13**). Intronic TEs facilitate the formation of hairpin structures in the mRNA [47]. The effect might be amplified in TE hotspots increasing the likelihood of circRNA occurrence. Finally, the enrichment of recently active SINE elements in long introns of GC-low genes (as seen in the flanking introns of circRNA parental genes) may be explained by the SINE-in-LINE integration bias and further supports the idea of a TE

Table 12: Highest and lowest circRNA frequencies reported | The Top 5 samples with the highest and lowest number of circRNAs reported are shown. *Frequencies were taken from circBase as of Apr 2017, [122]*

Organism	Sample	#circRNAs	Study
Human	Frontal cortex	38,983	Ashwal-Fluss *et al.*, 2014 [57]
Human	Occipital lobe	31,085	Ashwal-Fluss *et al.*, 2014 [57]
Human	K562	27,307	Salzman *et al.*, 2.013 [69]
Human	Diencephalon	24,632	Ashwal-Fluss *et al.*, 2014 [57]
Human	Parietal lobe	23,303	Ashwal-Fluss *et al.*, 2014 [57]
Mouse	Adult brain	537	Memczak *et al.*, 2013 [93]
Human	CD34+	528	Memczak *et al.*, 2013 [93]
Human	Neutrophil	417	Memczak *et al.*, 2013 [93]
Human	HEK293	239	Memczak *et al.*, 2013 [93]
Human	H9	103	Zhang *et al.*, 2013 [121]

Table 13: Fraction of coding and parental genes in isochores | Table provides an overview of the fraction (%) of coding and parental genes found in the different isochores.

Species	L1	L2	H1	H2	H3	type
Opossum	23.90	33.71	19.34	10.85	12.20	Coding gene
Mouse	3.66	23.41	31.71	24.32	16.90	Coding gene
Rat	2.74	20.46	30.08	27.2	19.52	Coding gene
Rhesus	8.05	24.91	20.62	17.43	28.99	Coding gene
Human	7.12	25.09	20.98	19.32	27.49	Coding gene
Opossum	11.49	3.81	1.07	0.48	0.23	Parental gene
Mouse	5.07	8.95	3.95	1.32	0.35	Parental gene
Rat	8.17	9.97	4.97	1.63	0.83	Parental gene
Rhesus	18.66	15.44	8.13	3.15	0.90	Parental gene
Human	20.82	20.46	9.54	3.40	1.11	Parental gene

hotspot, in which young TEs integrate into already existing TEs.

In summary, the high number of parental genes might be explained by the high number of coding genes that based on their structure (GC-low, long introns, many TEs) create a genomic context, in which the formation of circRNAs is facilitated. Consequently, circRNAs should rather be seen as a TE-driven stochastic property of a subgroup of genes than an actively maintained product.

3.2.2 Controversy 2 - CircRNAs can be highly expressed and must therefore be important

The majority of circRNAs is characterized by low expression levels. Standard RNA-seq experiments produce libraries with more than 40,000,000 reads of which in average 2-10 reads map on a single BSJ. To filter important circRNAs, several studies have calculated the linear-to-circular ratio (circ-ratio) and prioritized their analyses on circRNAs with high circ-ratios.

However, taking simply the ratio leads to a loss of information on the real expression levels. For instance, Guo *et al.* used 39 samples from different human cell lines and tissues to identify 7112 circRNAs, of which each circRNA was supported by at least two reads. In addition, the circRNA had to have in at least two samples a circ-ratio of >=0.1. However, when they inferred FPKMs, 60% of circRNAs had FPKMs below 1. To calculate the circ-ratio, they considered each transcript with an FPKM of at least 0.1. In a normal RNA-seq experiment, it is common practice to remove transcripts with less than 1 FPKM, because it is difficult to calculate correct expression levels for weakly expressed genes due to the high level of biological and technical noise. In addition, 29 out of 39 samples in the analysis of Guo *et al.* are from different cell lines and tissues, many of them without replicates. As discussed in **Chapter 1.1.3**, the stochastic effect is large in lowly expressed genes and even larger if one compares gene expression across a high number of different biological

samples. If expression levels are calculated for weakly expressed transcripts, it is important to have a high number of replicates. The circRNA candidates provided by Guo *et al.* based on high circ-ratios, are likely influenced by strong noise levels as the expression levels of the analyzed transcripts are low and have different biological backgrounds. One should therefore be careful when interpreting their results.

Unfortunately, Guo *et al.* is one of the most highly cited studies in the context of "strongly expressed circRNAs". Other studies have reported much lower numbers of circRNAs with a circ-ratio of >=0.1 (e.g. less than 10% of detected circRNAs in [65] and [69]). In my own data for instance, the median CPM for circRNAs in cerebellum ranges from 0.09-0.1 CPM, which corresponds roughly to one read per 10,000,000 reads. Shannon's equitability as calculated in **Chapter 2.3.2** suggests that circRNA expression levels in cerebellum are evenly distributed arguing that the majority of circRNAs have low read counts. When discussing circRNA abundance, one therefore needs to provide the circ-ratio together with the number of detected BSJ reads.

The low expression levels of circRNAs support the hypothesis that circRNAs are merely a by-product of TE integration and are driven by stochastic processes.

3.2.3 Controversy 3 - CircRNAs are abundant in neuronal tissues and are therefore important

CircRNAs are abundant and are differentially expressed between neuronal tissues and developmental stages of the brain [66, 67, 97, 123]. You *et al.* showed that in the murine brain, circRNAs are frequently found in genes with synaptic functions. In agreement with this, circRNAs are enriched in the neuropil - a region of the brain that consists of unmyelinated axons, dendrites and glial cells [123]. Similar findings were made by Veno *et al.* in the porcine brain [67], by Rybak-Wolf *et al.* in human brain tissues [66] and by Westholm *et al.* in the Drosophila brain [68]. Due to the high number of different circRNAs that were detected, all studies hypothesized that circRNAs play important roles in brain development and homeostasis. However, there are other possibilities that could explain the accumulation of circRNAs in neuronal tissues.

CircRNAs are very stable and until now, it is unknown if and how circRNA abundance is regulated in the cell. In cells with slow division rates, they might therefore accumulate throughout time. Starvation of fission yeast (Schizosaccharomyces pombe) for instance, leads to a decrease in cell proliferation and an increase in circRNA abundance [124]. In addition, Bachmayr-Heyda *et al.* described a negative correlation between the proliferation index of a cell and the frequency of

circRNAs: The number of circRNAs decreases in many tumor types when compared to the same normal tissue, and changes in expression profiles are more dynamic between cancer cells than for normal tissues [82]. Westholm *et al.* detected a positive correlation between the age of fruit flys and the number of circRNAs detected in their brains. They proposed that circRNAs could be used as a biomarker for aging [68]. Interestingly, aging has been associated with increased noise in gene expression [125]. Many cells in the adult brain are in a non-proliferative stage and together with the higher age of the adult brain, it might explain the enrichment of circRNAs in this organ. Further support for this hypothesis comes from the research field of somatic mosaicism.

Somatic mosaic mutations are defined as mutations that occur in only some cells of the soma in a single individual leading to a mixture of cells, which are distinguished by their DNA. Somatic mutations accumulate in postmitotic cells such as neurons and have mainly been associated with neurological diseases and aging. Interestingly, neuronal somatic mutations reflect damage that occurred during DNA transcription: Genes that are strongly expressed in a neuron accumulate more mutations [126]. Similar to somatic mutations, circRNAs may thus reflect the transcriptional state of a gene in postmitotic tissues.

The idea that circRNAs accumulate in postmitotic tissues can also explain why the number of identified circRNAs is lower in testis than in cerebellum. High transcriptome complexity is characteristic for testis, however, the tissue is characterized by a constant renewal of the spermatozoal repertoire. Thus, circRNAs do not have a lot of time to accumulate in a cell and in addition, the abundance of circRNAs likely depends on when it was measured during the spermatogenic cycle [21, 127].

The term structural plasticity describes the brain's ability to change its physical structure as a result of learning [128]. In a broader sense, each learning process could be seen as an environmental adaptation. However, to learn and to adapt efficiently, neurons need to be flexible. As discussed in **Chapter 1.1.3**, stochasticity is an important factor present in many cell populations and stochasticity allows for a robust response in changing environments [13]. Every day, neurons are exposed to different kinds of known and unknown information. Higher stochasticity in neuronal cells may thus be advantageous by increasing the chance that a neuron processes the information correctly. In such a case, increased stochasticity in transcription and splicing may correlate with the high number of different circRNAs in the brain. To my knowledge, there is currently no study, which has investigated a possible connection between structural plasticity and stochasticity in gene expression and splicing in the brain.

The enrichment of circRNAs in the brain could be explained by a combination of 1) low proliferation rates, which lead to the accumulation of circRNAs during aging and 2) heterogeneous neuronal cell populations that allow structural plasticity. In that case, the high abundance of circRNAs in neuronal tissues does not provide a strong argument for importance and functionality.

3.2.4 Controversy 4 - CircRNAs exhibit strong splicing signals and high phastCons scores and are therefore functional

CircRNA exons and splice sites have been associated with elevated conservation scores. Memczak *et al.* compared the phyloP scores at the third nucleotide position of 223 circRNAs present in human and mouse against a set of randomly selected control exons that were matched for the level of conservation observed in the first and second codon position. PhylopP scores were significantly larger at the third codon position of circRNAs than in the control exons [93]. Guo *et al.* used a similar approach and could confirm these findings. However, when they compared phyloP scores at the third codon position between circRNA exons and flanking exons, they did not find any significant difference [54]. You *et al.* compared phastCons scores between BSJs and linear junctions of the same gene and found significantly higher phastCons scores at BSJs [123]. In all three studies, elevated conservation scores were interpreted as a sign of functional relevance.

In my own data, circRNA exons have significantly higher phastCons scores than the remaining exons from the same gene. However, the effect is milder when the first and last exons are excluded (data not shown). In addition, I also found a stronger GC amplitude at circRNA exons when compared to other exons from the same gene (**Chapter 2.6.1**). At this point, I hypothesized that the elevated phastCons scores and amplitude may not be due to a conservation signal from the circRNA, but rather because circRNA exons are constitutive (included in all transcripts) and because the parental gene needs to adapt the splice site amplitude due to the repetitive landscape in the flanking introns. Selection may favor mutations that increase the amplitude to differentiate the exon from the intron, because GC-rich repeats can mask the GC amplitude at the exon boundary [3]. To analyze whether circRNA exons are constitutive, I compared the presence of circRNA exons to the presence of non-circRNA exons from the same gene in different transcripts expressed in liver, cerebellum and testis. CircRNA exons are more frequently used in multiple tissues when compared to the background dataset (data not shown). The inclusion level is not at 100%, which would define a constitutive exon. However, it is higher than expected suggesting that circRNA exons are frequently used in transcripts. I have tried to incorporate the ensembl annotation for constitutive exons. Yet,

the annotation quality varies a lot between opossum, mouse, rat, rhesus macaque and human and results are biased by the annotation quality. A more detailed analysis of exon inclusion frequencies across multiple species and tissues remains to be done.

Based on the obtained results, I would nevertheless argue that circRNA exons differ from their flanking exons, because they are frequently used in transcripts and need to be spliced correctly besides an increased number of TEs in the flanking introns. Therefore, elevated conservation scores are not driven by the functional importance of the circRNA, but by the importance of the exon in the linear counter-part.

3.2.5 Controversy 5 - CircRNAs are conserved between species and therefore, are likely functional

CircRNAs are frequently found in orthologous genes across different species. The observed overlap was used to propose the functional relevance of circRNAs [64, 66, 67].

In contrast to previous studies, the here-presented data indicates that the observed overlap of circRNAs between species is not caused by divergent evolution from a common ancestral circRNA, but instead by the independent occurrence of circRNAs in common genomic niches that are provided by the parental gene. The genomic niche is composed of a low GC content and long introns that allow for integration of transposable elements (discussed in **Chapter 3.2.1**). In addition, expression levels of the circRNAs are driven by the integration of recently active and species-specific SINE elements. In the subsequent, I would like to discuss several of my findings addressing the TE landscape and the overlap of circRNAs and hotspots between species to elucidate this point.

The overlap of circRNA hotspots across species

I classified circRNA loci as species-specific, lineage-specific, eutherian or therian depending on the number of species they occurred in (**Chapter 2.5.1**). CircRNA hotspots describe parental genes that produce at least two overlapping circRNAs. Interestingly, hotspots were often annotated as eutherian or therian, which means that they occur in the majority of species analyzed. In addition, parental genes of circRNA hotspots are found in the upper tail of the probability distribution that describes parental genes, indicating that the presence of a hotspot is correlated with a high probability of circRNA occurence (**Chapter 2.6.1**). Gene structure is often conserved between orthologous genes, which could explain why the same parental genes produce circRNAs in different species and why hotspots are often found in multiple species.

The overlap of circRNAs across species

In contrast to circRNA hotspots, circRNAs that share the same first and last exon are less frequently found across different species. CircRNAs are associated with species-specific and recently active TEs. The integration and fixation dynamics of TEs are specific for a species and can therefore explain why only a few circRNAs possess the same structure across different species despite the same parental gene.

The influence of species-specific SINE elements

The evolution of many SINE elements is closely connected to the evolution of the TE used for amplification and the host's defense mechanisms. Subsequently, many SINE elements exhibit species-specific amplification rates and patterns [32]. SINE elements are frequently found in introns of human and murine genes, and their presence is known to influence the splicing behavior and secondary mRNA structure of a gene [43, 47]. CircRNA flanking introns are enriched in repetitive elements. In human, many of these elements overlap with Alu elements and have been implicated in the hairpin formation that facilitates backsplicing [61, 64, 65]. A similar enrichment of SINE elements is also found in murine and porcine circRNA flanking introns [64, 67]. However, the exact distribution of different SINE families in flanking introns and their binding stability has not been studied.

I annotated reverse-complement alignments, which flank circRNAs, to describe the relative abundance of each TE and to identify preferential TE dimers. In all species, recently active TEs are enriched in flanking introns and in TE dimers. Furthermore, dominant circRNAs in a hotspot are enriched in recently active TEs suggesting that the presence of a circRNA is driven by active and species-specific SINE elements. Stable dimerization is more likely to occur between young SINE elements, because complementary is still high. With time, dimerization is less likely due to the degradation of the SINE element. CircRNA production is a dynamic process, in which the integration locus, the distance to the splice site and other SINE elements and the stability of the TE dimer play important roles. Because degradation and integration rates vary between species, this will lead to very species-specific circRNA landscapes.

TE integration and fixation dynamics could also explain the presence and depth of a circRNA hotspot. Whether a gene has the capability to produce a circRNA or not, depends on its structure. Hotspots can serve as an estimate for this capability. If it is high, the gene can produce many circRNAs (deep hotspot). In addition, gene structure is similar between orthologous genes causing the

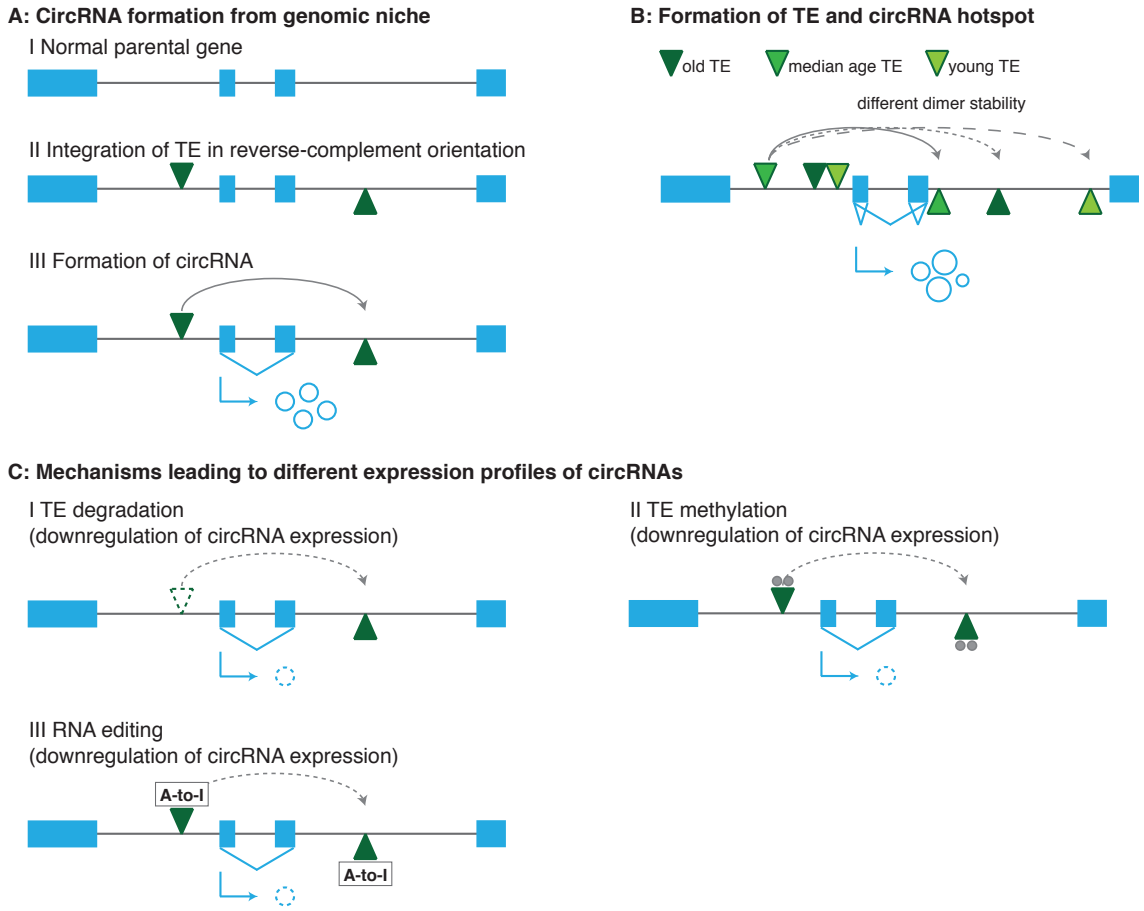


Figure 28: Co-evolution of circRNAs and TEs | A: CircRNA origin. The reverse-complement integration of TEs in introns can lead to the formation of a circRNA. **B:** Hotspot origin. Repetitive integration of TEs into each other can lead to the formation of an integration hotspot. This may facilitate the evolution of a circRNA hotspot with multiple circRNAs, depending on the different repeat dimers and hairpins that are formed. **C:** Differential circRNA expression. Degradation of the TE sequence (I), methylation by the host genome (II) or RNA editing (III) can lead to a decrease in circRNA production.

occurrence of hotspots in many species. Within each hotspot however, circRNA dynamics depend on the species-specific integration and fixation rates of TEs leading to a species-specific circRNA landscape. It is important to understand that it is the orthologous genomic region that leads to the independent occurrence of circRNAs across species. CircRNAs themselves are not conserved, only the underlying genomic locus is.

CircRNA expression profiles in light of SINE element dynamics

CircRNAs exhibit tissue-specific and species-specific expression profiles. It is likely that the observed differences are not explained by circRNA-specific regulatory mechanisms, but by the integration and

fixation dynamics connected to SINE elements. Recently integrated SINE elements can form stable dimers, which leads to stable hairpins and may cause an increase in circRNA abundance. With the degradation of the TE, dimer stability decreases leading to lower circRNA expression levels.

The secondary structure of the DNA or the mRNA influences the stability of a TE dimer. RNA-editing for instance, evolved as a mechanism to suppress TE amplification. A-to-I RNA editing is associated with intronic Alu elements to inhibit Alu dimers [46, 47]. In agreement with this observation, circRNA flanking introns are enriched in A-to-I editing and knockdown of the editing machinery leads to an increase of circRNA levels. Rybak-Wolf *et al.* suggested that A-to-I editing is a mechanism to control circRNA production [64]. However, changes in circRNA frequency are a secondary effect caused by the primary purpose of A-to-I editing during the inhibition of Alu amplification. Alu elements can be subject to other modification such as methylation, which interferes with amplification of the TE [129]. Similar to RNA editing, methylation can occur dynamically and in a tissue-specific manner causing the different spatio-temporal expression profiles pbserved for many circRNAs (**Figure 28C**). Recently, Aktas *et al.* provided evidence that the nuclear RNA helicase DHX9 can bind to inverted Alu elements, which are transcribed as part of the gene. Interestingly, the loss of DHX9 leads to an increase of circRNA abundance in parental genes. Aktas *et al.* suggested that DHX9 resolves TE-mediated secondary structures of the mRNA to avoid interference with post-transcriptional processes [130].

Dominant circRNAs are more often shared between species and in at least some cases, dominance is associated with a higher number of recently active repeats in the flanking introns. The formation of a strong TE hotspot takes time, which can explain why dominant circRNAs are more often shared. In a gene, there are probably only a few regions that allow for the formation of a TE hotspot. Given a similar structure of orthologous genes, this region will be present in all of them. There is only little flexibility of where in a gene the TE hotspot occurs leading to the formation of shared circRNAs from orthologous loci. Over time, circRNAs become dominant, because the TE hotspot becomes stronger with each new TE integration counter-acting the degradation of previous, underlying repeats.

Integration of a SINE element can lead to increased noise levels during transcription and splicing by interfering with regulatory elements. It is therefore likely, that circRNAs are a stochastic by-product of SINE integration into a gene. The observed dynamics in circRNA expression profiles reflect the co-evolution of TE amplification rates and TE silencing by the host genome.

3.3 The evolution of functional circRNAs

A small subset of circRNAs has been linked to a molecular function. But if as argued before, circRNAs are a by-product of stochastic gene expression and the presence of transposable elements, how can they then evolve a function?

The majority of transcriptional noise is probably neutral, but neutrality can be conditional. Conditionality in an organism refers to different cell types, stress situations or developmental transitions. Stochastic variations that are neutral in one cell type might be beneficial or deleterious in another. Dimitrii Polev developed the idea of an "in-service" mechanism of gene evolution, in which stochastic gene expression creates a broad landscape of transcriptional variation on which selection can act. For him, noisy gene expression is a compromise between the evolution of long-term benefits for an organism and its current needs. Because noise occurs broadly in every gene and tissue, the "in-service" mechanism can easily assess the quality of a novel variation in different cellular environments. If the variation interferes with fitness, selection will act on it [131]. An ubiquitously expressed gene for instance that produces a novel splicing variant, will allow the variant to be tested in many cellular environments and conditions. Because of its initial low abundance, the transcript's interactions with different cellular processes can be tested at little risk. The expression profile of a circRNA is linked to the parental gene by its transcription, splicing and the influence of the TE landscape. The interaction of these components creates a diverse circRNA landscape, in which circRNAs are exposed at different expression strength to different cellular environments.

CircRNAs can evolve a function, but the process is long and influenced by many components. In a first step, the circRNA needs to reach a critical threshold at which it can influence the organism's fitness. Reaching the threshold is for instance influenced by the mitotic status of the cell it is expressed in. Because highly proliferative cells dilute the circRNA concentration, the threshold might only be reached in postmitotic cells such as neurons (**Figure 29**).

Novel transcripts can evolve by a process known as exaptation, in which a novel genomic feature is not build by natural selection as adaptation for its current function, but is coopted from an already existing structure that performs a different function (e.g. feathers were initially thermoregulatory, but later coopted for flight) [132]. The concept has widely been accepted to explain the development of novel exons from repetitive elements or the evolution of miRNAs from introns. Exaptation can accelerate evolution, because certain sequence structures and regulatory elements are already in place. More importantly, they can lead to the parallel evolution of similar features in the same

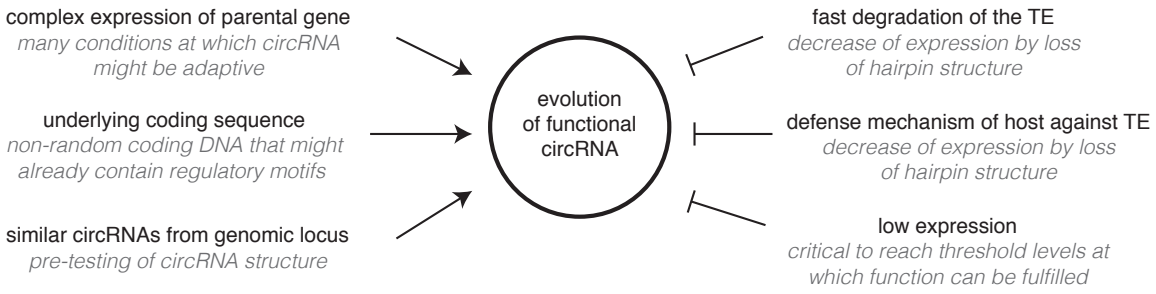


Figure 29: Factors influencing the evolution of a functional circRNA | Different factors can influence the evolution of a functional circRNA. Parental genes that are broadly expressed increase the chance of the circRNA being present at a beneficial condition. The underlying coding sequence might already contain regulatory motifs due to its coding nature, and circRNAs with similar structure might have been pre-tested by the locus. The degradation of the TE and defense mechanisms of the host genome may reduce the likelihood of a hairpin structure and thus the formation of a circRNA. The low expression levels of a circRNA make it difficult to reach a threshold at which the functional influence is strong.

genomic context. Several miRNAs for instance, evolved independently from each other in the same relative genomic position to the *hox8* gene. The parallel evolution is explained by a strong selective pressure for the evolution of a miRNA in this region to regulate the posttranscriptional control of the *hox8* gene [133]. Two enhancer of the proopiomelanocortin gene (*POMC*) provide another example of exaptation. *POMC* is expressed in a group of neurons in the hypothalamus. Expression is regulated by two enhancer that evolved independently from an LTR and a SINE element into neuronal-specific enhancer [134]. The sequence composition of a circRNA is not random, because it is derived from coding DNA. Different DNA motifs may already exist, which can be adapted for a novel function independent of the coding sequence they are coming from. Similar to the previously described miRNAs, functional circRNAs may evolve by exaptation in the same genomic context of different species (**Figure 29**).

The evolution of a circRNA is a dynamic process, in which the integration of new TEs, their complementarity and proximity to each other and the next splice site determine which circRNA will be produced. These dynamics are reflected in circRNA hotspots, in which multiple circRNAs with distinct, but overlapping structure and expression co-exist. CircRNAs from a hotspot provide therefore a set of structurally similar RNAs, from which evolution can chose the most promising. However, the degradation of TEs and defense mechanisms of the host genome will interfere with circRNA production. In many cases, circRNAs may only occur at low levels and for a short period of time, which will make it unlikely for them to evolve functionality (**Figure 29**).

I have described several processes, which can influence the evolution of a functional circRNA.

However, all of them take place in a single individual. In the next steps, processes such as random genetic drift and different selective pressures will act on the frequency of a circRNA allele in a population and determine whether the circRNA allele remains in the population and eventually becomes fixed.

In summary, there are many processes that can either accelerate or inhibit the evolution of a functional circRNA. Because they are frequent and occur in many cell types and developmental time points of an organism, it is not impossible that some beneficial variants may evolve. However, the evolution of a function takes time and is a complex process, during which the majority of circRNA variants will not succeed.

Given a set of predicted circRNAs, how can one prioritize the list for further experiments? If circRNA expression is stochastic, then functional circRNAs should be the least stochastic. If enough replicates are provided, it should be possible to look at the variation across replicates and to select candidates with a low variance. Genes with higher noise levels are more likely to produce circRNAs as a stochastic side-product. Low-variation circRNAs produced by genes with robust expression profiles and reduced noise levels could therefore be the most likely candidates for follow-up studies. Expression levels of the circRNA may help to decide, but should be interpreted in light of the mitotic state of the tissue. Is it still proliferating and high circRNA abundance is therefore unexpected, or is the tissue postmitotic and may simply have accumulated circRNAs through time? When analyzing circRNA candidates for a function in cell homeostasis or disease, it is therefore important to keep the different mechanisms that can create dynamic circRNA expression profiles in mind.

3.4 CircRNAs as disease biomarkers

CircRNAs are very stable and their potential to act as biomarker for human diseases was quickly discovered [90, 91]. The assumption that circRNAs are mainly transcriptional noise does not influence this finding. Disturbance of the local noise level might be an important factor for many diseases. The analysis of circRNAs in this context could link the phenotype to potential problems in the transcriptional or splicing control of the parental gene, novel TE integrations or problems with the RNA editing or methylation machinery. A careful analysis of circRNAs that are upregulated or downregulated in a disease phenotype might thus help to detect the disease cause.

Furthermore, the high stability of circRNAs could help to detect diseases already at early onset, because disturbances of the local noise level might be the first symptomes that can be observed.

Often, patient samples are subject to fast RNA degradation rates, which can impair and weaken subsequent analyses. CircRNAs in contrast are more stable, and may thus not only facilitate sample analysis, but maybe lead to higher quality results.

3.5 Final summary and outlook

Throughout this thesis, I have tried to reason for the stochastic nature of circRNAs. The main arguments I have made can be summarized as follows:

1. Parental genes exhibit a similar structure: They are GC-poor, have large introns, a high splicing capability and are characterized by a high TE integration rate. The co-occurrence of these components creates a genomic niche that predisposes parental genes for the production of circRNAs.

2. Genomic niches are similar between orthologous genes, because they often have a conserved structure and genomic location. CircRNA hotspots that occur in orthologous parental genes indicate the conservation of this niche.

3. Genomic niches are subject to the integration of species-specific transposable elements. Independent TE integration events have led to the parallel emergence of circRNAs across species. CircRNAs that are shared between species did not derive from a common ancestral circRNA.

4. The emergence of novel circRNAs is a dynamic process in which circRNA expression levels and structures are a function of the number of integrated TEs, their pairing stability and distance to each other. CircRNAs that are dominant in a hotspot are connected to recent integration events.

I have provided first evidence for the stochastic nature of circRNAs. But to support this hypothesis several additional analyses should be done, either as part of this project or independently.

Evolutionary contrasts

If the presence of a circRNA relies on the genomic context, then orthologous genes that differ in their structure from each other could further differentiate the importance of the different components such as the GC content or repeat landscape. Given for example a circRNA that occurs in four out of five species, can the lack of the circRNA in the fifth species be explained by a higher GC content, smaller introns or a lack of transposable elements in the orthologous gene?

Stochasticity and structure of the parental gene

If the stochasticity of the parental gene in transcription and splicing and its structure are connected to the presence of a circRNA, then it should be possible to develop a scoring system to assess the probability of a gene to produce circRNAs. Such a scoring system could be used to predict the likelihood of a parental gene without the need of doing RNA-seq experiments.

Improvement of linear models and alternatives

I have used multiple linear regression to model my circRNA data based on a combination of different structural and functional predictors. In the here-presented results, I have fitted the models on the whole dataset without separating the data into training and prediction datasets, something that should be improved. Furthermore, I have used a set containing all coding genes, while the circRNA detection was restricted to three tissues. Reducing the set of coding genes further to genes that are expressed in these three tissues might improve the model. Furthermore, the linear model could be trained on liver, cerebellum and testis genes and subsequently be used to predict circRNAs in additional tissues.

In multiple linear regression the response variable is modeled as a linear combination of different predictive variables. But linear models risk to fail or to give low prediction accuracy if sample size is small and if different predictors are correlated with each other. In my analyses, many of the predictors are correlated with each other (GC content, gene length, number of TEs, etc.). Although the tests, which were performed to assess collinearity between the predictors suggested that the observed interactions can be neglected, it might be better to use methods based on machine learning approaches.

Interaction between TE integration and circRNA expression

Transposable elements play an important role in the production of a circRNA. To approximate their influence, I have looked at the relative abundance of TEs in different contexts (e.g. parental genes vs. non-parental genes, flanking introns vs. background introns or dominant vs. non-dominant circRNAs). The approximation could be improved by reconstructing the composition of TE hotspots in the flanking introns (which TEs, size, distance to splice site) and by understanding its evolution (which TEs integrated into which TE, how strong are the degradation rates). The strategy used by Levy *et al.* to reconstruct intergenic TE hotspots could be used here [28].

Furthermore, it would be of great interest to understand how the interaction between recent TE integrations, circRNAs and host defense mechanisms evolves. Different mechanisms that are associated with TE silencing such as RNA-editing or methylation should be further investigated in this light.

The role of stochasticity in the evolution of tissues

Independent of circRNAs, it might be from great interest to investigate transcriptional noise levels between tissues to understand potential correlations between the evolution and function of a tissue in light of its stochastic properties.

Finally, this study could also be seen as a gentle reminder that it is important to analyze one's data carefully in the context of different evolutionary and genetic concepts. These concepts need to be well understood before applying them to large datasets and novel research directions to avoid misinterpretations and oversimplification. Unfortunately, in a world of ever-larger research projects, interdisciplinarity, high-publishing pressure and a teach-it-yourself attitude, scientists often struggle to develop and to phrase their research hypotheses correctly. Active and open participation in different conferences and courses is therefore key to broaden their understanding and the interpretation of their own data!

4 Methods

4.1 Programs and working environments

Table 14: Overview of external programs

Program	Version
Blast	2.2.29+
BEDTools	2.17.0
Bowtie2	2.1.0
Cufflinks	2.1.1
FastQC	0.10.1
Latex	MacTEX-2013 installation package
Mcl	14.137
R	3.0 and 3.1
Ruby	2.0 and 2.1
SAMTools	0.1.19
TopHat2	2.0.11
ViennaRNA	2.1.8

4.2 Library preparation and sequencing

**Division of tasks: Libraries were performed under supervision of Peggy Janich in the laboratory of David Gatfield at the Center of Integrative Genomics of the University of Lausanne.*

We used 6 µg of RNA/tissue for the library preparation. Of these 6 µg, 5 µg were treated with RNase R for 1 h at 37°C to degrade linear RNAs prior library preparation. Libraries were prepared with the Illumina TruSeq Stranded Total RNA with Ribo-Zero Gold according to the protocol with the following exceptions to select larger DNA fragments: 1) Instead of the recommended 8 min at 68°C for fragmentation, we incubated samples for only 4 min at 68°C to increase the fragment size. 2) In the final PCR clean-up after enrichment of the DNA fragments, we changed the 1:1 ratio of DNA to AMPure XP Beads to a 0.7:1 ratio to select for binding of larger fragments. Libraries were analyzed on the fragment analyzer for their quality and sequenced with the Illumina HiSeq 2500 platform.

Table 15: RNase R treatment | Samples were treated for 1 h at 37°C according to the presented master mix. RNA was purified with the Zymo RNA purification research kit and RNA recovered in 10 µl of water.

Component	Concentration
RNA	5 µg
RNase R reaction buffer (10x)	2 µl
RNase R	1 µl
Water	up to final volume of 20 µl

4.3 Identification and quantification of circRNAs

4.3.1 Mapping of RNA-seq data

I downloaded the ensembl annotations for opossum (monDom5), mouse (mm10), rat (rn5), rhesus macaque (rheMac2) and human (hg38) from `http://www.ensembl.org/info/data/ftp/index.html` to build transcriptome indexes for mapping with TopHat2. TopHat2 was run with default settings and the *–mate-inner-dist* and *–mate-std-dev* options set to 50 and 200 respectively. The mate-inner-distance parameter was estimated based on the fragment analyzer report.

Table 16: Ensembl genome versions and annotation files for each species

Species	Genome	Annotation
Opossum	monDom5	ensembl release 75, feb 2014
Mouse	mm10	ensembl release 75, feb 2014
Rat	rn5	ensembl release 75, feb 2014
Rhesus macaque	rheMac2	ensembl release 77, oct 2014
Human	hg38	ensembl release 77, oct 2014

4.3.2 Analysis of unmapped reads

I developed a custom pipeline to detect circRNAs, which performs the following steps: Unmapped reads with a phred quality value of at least 25 are used to generate 20 bp anchor pairs from the terminal 3' and 5'-ends of the read. Anchors are remapped with bowtie2 on the reference genome. Mapped anchor pairs are filtered for 1) being on the same chromosome, 2) being on the same strand and 3) for having a genomic mapping distance to each other of a maximum of 100 kb. Next, anchors are extended upstream and downstream of their mapping locus. They are kept if pairs are extendable to the full read length. During this procedure a maximum of two mismatches is allowed. Next, all unpaired reads are selected from the accepted_hits.bam file generated by TopHat2 (singletons) and assessed for whether the mate read (second read of a paired-end sequencing read) of the anchor pair mapped between the backsplice coordinates. All anchor pairs for which 1) the mate did not map between the genomic backsplice coordinates, 2) the mate mapped to another backsplice junction or 3) the extension procedure could not reveal a clear breakpoint are removed. Based on the remaining candidates, a backsplice index is build with bowtie2 and all reads are remapped on this index to increase the read coverage by detecting reads that cover the BSJ with less than 20 bp, but at least 8 bp. Candidate reads that were used to build the backsplice index and now mapped to another backsplice junction are removed. Upon this procedure, the pipeline provides a first list of backsplice

junctions.

The set of scripts, which performs the identification of putative BSJs, as well as a short description of how to run the pipeline are reposited at GitHub: `https://github.com/Frenzchen/ncSplice_circRNAdetection`.

4.3.3 Trimming of overlapping reads

Due to the small DNA repeats, some reads are extendable to more than the original read length. Therefore, overlapping reads were trimmed based on a set of canonical and non-canonical splice sites. For the donor site GT, GC, AT, CT were used and for the acceptor splice site AG and AC. The trimming is part of previously described custom pipeline and the step will be performed automatically if the scripts are run.

4.3.4 Calculation of CPM value

CPM (counts per million) values for BSJs were calculated for each tissue as follows:

$$counts = mean(counts_{rep1}, counts_{rep2}, counts_{rep3})$$
$$mappedReads = mean(mappedReads_{rep1}, mappedReads_{rep2}, totalMappedReads_{rep3})$$
$$CPM = counts * 10^6 / totalMappedReads$$

4.3.5 Filtering of candidates based on CPM enrichment

To distinguish putative BSJs from the technical and biological noise background, the enrichment of the previously defined junctions in RNase R treated samples was calculated. The enrichment was defined as CPM increase in RNase R treated versus untreated samples:

$$enrichment = CPM_{RNaseR} / CPM_{untreated}$$

Candidates with a log2-enrichment of smaller 1.5 were removed.

4.3.6 Manual filtering steps

I observed a couple of genomic loci, which were highly enriched in reads for putative BSJs. Manual inspection in the UCSC genome browser indicated that these loci are highly repetitive. The detected BSJs from these regions do probably not reflect BSJs, but instead issues in the mapping procedure.

I removed these candidates manually. All following analyses were conducted with the circRNA candidates that remained until now.

4.3.7 Calculation of Shannon diversity index and Shannon's equitability

The Shannon diversity index H and Shannon's equitability EH were calculated for circRNAs in each tissue and species. CPM values served as approximation for the abundance of individual circRNAs, p reflects the probability of a circRNA to occur in a given tissue and species and n equals the total number of circRNAs.

$$p = CPM_{circRNA}/CPM_{total}$$
$$H = -\sum p * log2(p)$$
$$EH = H/log2(n)$$

4.3.8 Reconstruction of circRNA isoforms

To reconstruct the exon structure of circRNA transcripts in each tissue, I made use of the junction enrichment in RNase R treated samples. To normalize junction reads across libraries, I calculated size factors based on the geometric mean of common junctions in untreated and treated samples,

$$geometric\ mean = product(x)^{1/length(x)}$$
$$size\ factor = median(x/geom.mean)$$

with x being a vector containing the number of reads per junction. I then compared read coverage for junctions outside and inside the BSJ for each gene. I used the log2-change of junctions outside the backsplice junction to construct the expected background distribution of change in junction coverage upon RNase R treatment. I then compared the observed coverage change of junctions inside the backsplice to the expected change in the background distribution and assigned junctions with a log2-change outside the 90% confidence interval as circRNA junction. I decided to choose a loose cut-off here, because involved junctions can show a decrease in coverage if their linear isoform was present at high levels before (degradation levels of linear isoforms do not correlate with the enrichment levels of circRNAs). Next, I reconstructed a splicing graph for each circRNA candidate, in which network nodes are exons connected by splice junctions (edges). Connections between nodes are weighted by

the coverage in the RNase R treated samples. The resulting network graph is directed (because of the known circRNA start and stop coordinates), acyclic (because splicing always proceeds in one direction), weighted and relatively small. I used a simple breadth-first-search algorithm to traverse the graph and to define the strength for each possible isoform by its mean coverage. For the further analyses, I considered only the strongest isoform.

4.4 Reconstruction and expression quantification of linear mRNAs

I reconstructed linear isoforms based on the pipeline provided by Trapnell *et al.* (Cufflinks + Cuffcompare + Cuffnorm) [135]. Expression levels were quantified based on fragments per million mapped reads (FPKM). Cufflinks was run per tissue and annotation files were merged across tissues with Cuffcompare. Expression was quantified with Cuffnorm based on the merged annotation file. All programs were run with default settings. FPKM values were normalized across species and tissues using a median scaling approach as described in [22].

4.5 In-vitro validation of candidates

**Division of tasks: qPCRs were performed under supervision of Peggy Janich in the laboratory of David Gatfield at the Center of Integrative Genomics of the University of Lausanne.*

4.5.1 cDNA synthesis

We prepared cDNA for total RNA (primer: random hexamer, reaction 1), poly(A)-selected RNA (primer: Oligo(dT), reaction 2) and RNase R treated total RNA (primer: random hexamer, reaction 3). As input we used 10 µg of RNase R treated RNA and 2 µg of total RNA. Primer were added according to **Table 17** in addition with dNTPs to a total volume of 12 µl. The mixture was incubated for 5 min at 65°C. To each mix, we added 4.5 µl 5x first-strand buffer, 2 µl 0.1M DTT and 1 µl Rnase inhibitor (promega). Oligo(dT) reactions were incubated for 2 min at 42°C and random-hexamer reactions for 2 min at 25°C. 1 µl of SuperscriptII reverse transcriptase was added. Transcription was performed with following programs:

- Oligo(dT): 60 min @ 42°C, 15 min @ 70°C, cooling to 4°C
- Random hexamer: 10 min @ 25°C, 50 min @ 42°C, 15 min at 70°C, cooling to 4°C

For the real-time (RT) PCR and qPCR, all samples were diluted to a final concentration of 5 ng/µl.

Table 17: cDNA synthesis

Component	Concentration	Reaction
Total RNA	2 µl + 0.5 µl random hexamers	1
Total RNA	2 µl + 1 µl Oligo(dT)	2
RNase R treated RNA	10 µl + 0.5 µl random hexamers	3
dNTPs	1 µl	1,2,3
water	up to 12 µl	1,2,3
RT	1 µl	1,2,3

4.5.2 Primer validation and qPCR

To validate the circRNA candidates, we designed divergent primer pairs with one primer falling on the backsplice junction itself and the second primer mapping to a downstream exon. We tested all primer pairs on cDNA from total RNA (reaction 1), poly(A)-selected RNA (reaction 2) and RNase R treated RNA (reaction 3) according to following qPCR protocols:

Table 18: qPCR master mix

Component	Concentration
cDNA	5 ng/µl
fwd primer	0.3 µl
rev primer	0.3 µl
SYBR green Master Mix	10 µl

Table 19: qPCR program At cycle 42, a melting curve for values between 45-90.1°C is generated by reading the plate at steps of 1°C (each step 5 sec).

Cycle number	Denaturing	Annealing	Extension	Read plate?
1	95°C, 15 min	-	-	no
2-40	95°C, 15 sec	57°C, 15 sec	72°C, 15 sec	yes
41	-	-	72°C, 7 min	yes
42	melting curve			
43	10°C, 20 sec, end			

4.6 CircRNA overlap between species

4.6.1 Identification of shared circRNA

Shared circRNAs were defined on three different levels depending on whether the parental gene, the circRNA locus in the gene or the start/stop-exons overlapped between species. One-to-one (1:1) therian orthologous genes were defined between opossum, mouse, rat, rhesus macaque and human using the ensembl orthology annotation (confidence intervals 0 and 1, restricted to clear one-to-one orthologs). The same procedure was performed to receive the 1:1 orthologous genes for the

eutherians (mouse, rat, rhesus macaque, human), for rodents (mouse, rat) and primates (rhesus macaque, human). To calculate the expected value of parental gene overlap based on 1:1 orthologs, the probability p of having an event i (e.g. a 1:1 ortholog in opossum, mouse, rat, rhesus and human parental genes) was calculated and multiplied with the number of trials (N), latter corresponding to the observed frequency of the event.

$$expected\ value = N * \sum p(i)$$

The circRNA overlap between species was assessed by counting the number of 1:1-orthologous parental genes between the five species. The analysis was restricted to protein-coding genes. To identify shared circRNA loci, all circRNA exon coordinates from a given gene were collapsed into a single transcript using the *bedtools merge* option from the BEDTools toolset with default options. Next, I used liftOver to compare exons from the collapsed transcript between species. The minimal ratio of bases that need to overlap for each exon was set to 0.5 (*-minMatch=0.5*). I defined collapsed transcripts as overlapping between different species if they shared at least one exon independent of the exon length. To identify circRNAs that overlapped with the same first and last exon, I lifted their coordinates between species (*liftOver, -minMatch=0.5*). CircRNAs were defined as overlapping if both exons were found as first and last exons in a circRNA of another species.

4.6.2 Identification of circRNA clusters for species overlap

I categorized overlapping circRNAs in the following groups: Species-specific, rodent, primate, eutherian and therian circRNAs. To be part of the rodent or primate lineage, the circRNA has to be expressed in both species of the lineage. To be part of the eutherian lineage, the circRNA has to be expressed in three out of mouse, rat, rhesus macaque and human. To be part of the therian lineage, the circRNA needs to be expressed in opossum and in three out of the four other species. Species-specific circRNAs are either present in one species or do not match any of the other four categories. To define the different groups, I used the cluster algorithm MCL [136, 137]. MCL is frequently used to reconstruct orthology clusters based on blast results. It requires input in *abc* format, in which a corresponds to event a, b to event b and a numeric value c that provides information on the connection strength between event a and b (e.g. blast p-value). If no p-values are available as in my analyses, the connection strength can be set to 1. MCL was run with a cluster granularity of 2 (option *-I*).

$ mcxload -abc species.abc –stream-mirror -o species.mci -write-tab species.tab

$ mcl species.mci -I 2

$ mcxdump -icl out.species.mci.I20 -tabr species.tab -o dump.species.mci.I20

4.6.3 PhastCons scores

PhastCons scores for exons of parental genes were calculated using the conservation scores provided by the UCSC genome browser (mouse: phastCons scores based on alignment for 60 placental genomes, rat: phastCons scores based on alignment for 13 vertebrate genomes, human: phastCons scores based on alignment for 99 vertebrate genomes). Mean phastCons scores were calculated for each exon in a coding gene. Next, the grand mean of circRNA-contained exons was divided by the grand mean of circRNA-outside exons on a gene-wise level (phastcons-ratio). A one-tailed Wilcoxon rank sum test was used to compare the log2-transformed phastcons-ratio to the expected value μ=0.

4.6.4 Expression clustering of circRNA and parental gene expression

To cluster therian circRNAs, I calculated the circular-to-linear ratio based on CPMs for BSJ reads and spliced linear reads covering the first and last exon of each circRNA. Based on the ratio, I calculated Spearman's-rank correlation coefficients between pairs of samples for all species and tissues. I then calculated the euclidian distance between samples with the default functions provided in the R environment (cor() and heatmap.2() from the *gplots* library). A similar approach was used to cluster expression levels of parental genes. However, FPKM values as calculated in **Chapter 4.4** were used instead of CPM-ratios.

4.7 Parental gene analysis

4.7.1 GC content of exons and intron

The ensembl annotation for each species was used to receive the different known transcripts in each coding gene. Transcripts were collapsed per-gene to define the exonic and intronic parts. Introns and exons were distinguished by their relative position to the circRNA (flanking, inside or outside). The GC content was calculated based on the genomic DNA sequence. On a per-gene level, the median GC content for each exon and intron type was used for the further analyses. Differences

between the GC content were assessed with a one-tailed Mann-Whitney U test.

4.7.2 GC amplitude

The ensembl annotation for each species was used to receive the different known transcripts in each coding gene. For each splice site, the GC amplitude was calculated using the last 250 intronic bp and the first 50 exonic bp. Splice sites were distinguished by their relative position to the circRNA (flanking, inside or outside). A one-tailed and paired Mann-Whitney U test was used to assess the difference in GC amplitude between circRNA-related splice sites and others.

4.7.3 Gene self-complementarity

The genomic sequence of each coding gene (first to last exon) was aligned against itself in sense and antisense orientation using megaBLAST with the following call:

$ blastn -query seq.fa -subject seq.fa -task dc-megablast -word_ size 12 -outfmt "6 qseqid qstart qend sseqid sstart send sstrand length pident nident mismatch bitscore evalue" > blast.out

The resulting alignments were filtered for being purely intronic (no overlap with any exon). The fraction of self-complementarity was calculated as summed length of all alignments in a gene divided by its length (first to last exon).

4.7.4 Frequency of transposable elements

The RepeatMasker annotation was used to identify the total number of intronic repeats in each coding gene. Repeat coordinates were intersected with all introns using BEDTools. A repeat was counted if its complete sequence was found in the intron (*bedtools merge -f 1*).

4.7.5 GO annotation

GO enrichment analysis was performed with GOrilla using parental genes against a background dataset of expressed coding genes in L1, L2 and H1 [138]. GO terms that were identified for mouse, rat and human were merged to identify common terms.

4.7.6 Integration of external studies

Replication time

Values for the replication time were used as provided in Koren *et al.* [111]. Coordinates of the different replication domains were intersected with the coordinates of coding genes using BEDtools (*bedtools merge -f 1*). The mean replictaion time of each gene was used for subsequent analyses.

Gene expression steady-state levels

Gene expression steady-state levels and decay rates were used as provided in Table S1 of Pai *et al.* [112].

GHIS

Genome-wide haploinsufficiency scores for each gene were used as provided by Steinberg *et al.* in Supplementary Table S2 [113].

4.8 Linear regression

4.8.1 Generalized linear models

All linear models were developed in the R environment. The different predictors used for each model are listed in **Table 6** and **Table 9**. The presence of multicollineraity between predictors was assessed using the vif() function from the R package *car* to calculate the variance inflation factor (VIF). Predictors were removed from the model if the VIF was greater than 5. Predictors that correlated with each, but still explained independent parts of the data were allowed to interact in the model. Predictors were scaled to be able to compare them with each other using the scale() function as provided in the R environment.

Each model was fitted on the complete dataset using all predictors. By using a step-down approach, individual predictors were removed manually starting from the least significant. The distribution of residuals for the final model was assessed for normality using the R function qqnorm() and the function residualPlots() from the package *statmod*. The final model was compared to the null model using the pchisq() function from R. Confidence intervals for each predictor were calculated using the confint() function.

To assess the presence and absence of parental genes, hotspot genes and the overlap between species (shared or species-specific), a binomial model was chosen, because it can reflect the presence

or absence of the response variable (0 probability for absence and 1 probability for presence). To address the depth of each hotspot, the poisson family of models was used, because the depth can be assessed on a count base. The initial model calls are listed in the following:

Presence or absence of parental gene, parental hotspot genes, shared and species-specific circRNA loci:

$$\begin{aligned} glm(data,\ response \sim\ & scale(GC\ content) * scale(genomic\ length) \\ & + scale(transcript\ count) \\ & + scale(ss.repeats/genomic\ length) \\ & + scale(as.repeats/genomic\ length) \\ & + scale(gene\ FPKM) \\ & + scale(tsi) \\ & + scale(phastcons) \\ & + factor(circRNA\ overlap),\ family = "binomial") \end{aligned}$$

Depth of the hotspot:

$$\begin{aligned} glm(data,\ depth \sim\ & scale(GC\ content) \\ & + scale(hotspot\ length) \\ & + scale(number\ of\ repeats) \\ & + scale(\Delta G) \\ & + factor(circRNA\ overlap),\ family = "poisson") \end{aligned}$$

4.8.2 Mixed linear models

Mixed linear models were fitted using the lmer() function from the package *lme4*. P-values were calculated by loading the package *lmerTest* in parallel. In general, models were fitted following the same procedure as described in the previous chapter. However, to address which circRNA is the dominant circRNA in each hotspot, the hotspot itself was added as background effect. The analysis was restricted to hotspots with 2-5 circRNAs. The dominance of each circRNA was approximated by rank-transforming its expression (rank 1 = most highly expressed, rank 2 = second highest ex-

pression etc.). Rank-transforming only leads to a situation in which the strongest circRNA has the lowest rank. Therefore, the reciprocal of each rank (1/rank 1, 1/rank 2 etc.) was used to keep the positive relationship between rank and increased probability.

CircRNA dominance in hotspot:

$$\begin{aligned} lmer(data, 1/rank \sim\ & scale(genomic\ length) \\ & + scale(number\ of\ repeats) \\ & + scale(acceptor\ amplitude) \\ & + scale(donor\ amplitude) \\ & + scale(mean.distance\ repeat\ to\ splice\ site) \\ & + scale(GC\ content\ intron) \\ & + factor(overlap\ circRNA) \\ & + (1|hotspot))\ \#background\ effect \end{aligned}$$

4.9 Repeat analyses

4.9.1 Generation of length- and GC-matched background dataset

Flanking introns were grouped into a matrix of i columns and j rows representing different genomic lengths and GC content. i and j were calculated in the following way:

$$i = seq(from = quantile(GCcontent,\ 0.05),\ to = quantile(GCcontent,\ 0.95),\ by = 0.01)$$
$$j = seq(from = quantile(length,\ 0.05),\ to = quantile(length,\ 0.95),\ by = 1000)$$

Flanking introns were sorted into the matrix based on their GC content and length. A second matrix with the same properties was created containing all introns of coding genes. From the latter, a submatrix was sampled with the same length and GC distribution as the matrix for flanking introns. The length distribution and GC distribution of the sampled introns reflect the distributions for the flanking introns as assessed by a Fisher's t Test that was non-significant.

4.9.2 Repeat enrichment in flanking introns

The total number of individual transposable elements was determined by intersection of the flanking and background introns with the RepeatMasker annotation (*bedtools merge -f 1*). The enrichment of flanking introns was assessed by a one-tailed Mann-Whitney U test.

4.9.3 Identification of repeat dimers and their binding stability

The complementary regions that were defined with megaBLAST as previously described were intersected with the coordinates of individual repeats from the RepeatMasker annotation. To be counted, a repeat had to overlap with at least 50% of its length with the region of complementarity (*bedtools merge -f 0.5*).

The free energy of the secondary structure of individual TE dimers was calculated with the RNAcofold function from the ViennaRNA Package:

$ RNAcofold -a -d2 < dimerSequence.fa

5 References

References

[1] E Melamud and J Moult. "Stochastic noise in splicing machinery". In: *Nucleic acids research* (2009).

[2] Giorgio Bernardi. "The genome: an isochore ensemble and its evolution." In: *Annals of the New York Academy of Sciences* (2012).

[3] M Amit et al. "Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition". In: *Cell reports* (2012).

[4] Zakharia M Frenkel, Thomas Bettecken, and Edward N Trifonov. "Nucleosome DNA sequence structure of isochores." In: *BMC genomics* (2011).

[5] Annalisa Varriale and Giorgio Bernardi. "Distribution of DNA methylation, CpGs, and CpG islands in human isochores." In: *Genomics* (2010).

[6] Alexander E Vinogradov. "Noncoding DNA, isochores and gene expression: nucleosome formation potential." In: *Nucleic acids research* (2005).

[7] A E Vinogradov. "Isochores and tissue-specificity". In: *Nucleic acids research* (2003).

[8] M Costantini. "An isochore map of human chromosomes". In: *Genome research* (2006).

[9] N Galtier and D Mouchiroud. "Isochore evolution in mammals: a human-like ancestral structure." In: *Genetics* (1998).

[10] Adam Eyre-Walker and Laurence D Hurst. "OPINION: The evolution of isochores". In: *Nature reviews. Genetics* (2001).

[11] S M Berget. "Exon recognition in vertebrate splicing." In: *The Journal of biological chemistry* (1995).

[12] Christopher V Rao, Denise M Wolf, and Adam P Arkin. "Control, exploitation and tolerance of intracellular noise." In: *Nature* (2002).

[13] Mads Kaern et al. "Stochasticity in gene expression: from theories to phenotypes." In: *Nature reviews. Genetics* (2005).

[14] Roy D Dar et al. "Transcriptional burst frequency and burst size are equally modulated across the human genome." In: *Proceedings of the National Academy of Sciences* (2012).

[15] Raheleh Salari et al. "Teasing apart translational and transcriptional components of stochastic variations in eukaryotic gene expression." In: *PLoS computational biology* (2012).

[16] Iksoo Huh et al. "DNA methylation and transcriptional noise." In: *Epigenetics & chromatin* (2013).

[17] Alfonso Martinez Arias and Penelope Hayward. "Filtering transcriptional noise during development: concepts and mechanisms." In: *Nature reviews. Genetics* (2006).

[18] Joseph K Pickrell et al. "Noisy splicing drives mRNA isoform diversity in human cells." In: *PLoS genetics* (2010).

[19] Eugene Melamud and John Moult. "Structural implication of splicing stochastics." In: *Nucleic acids research* (2009).

[20] Hadas Keren, Galit Lev-Maor, and Gil Ast. "Alternative splicing and evolution: diversification, exon definition and function." In: *Nature reviews. Genetics* (2010).

[21] Magali Soumillon et al. "Cellular source and mechanisms of high transcriptome complexity in the mammalian testis." In: *Cell reports* (2013).

[22] David Brawand et al. “The evolution of gene expression levels in mammalian organs.” In: *Nature* (2011).

[23] Anamaria Necsulea and Henrik Kaessmann. “Evolutionary dynamics of coding and non-coding transcriptomes.” In: *Nature reviews. Genetics* (2014).

[24] Nuno L Barbosa-Morais et al. “The evolutionary landscape of alternative splicing in vertebrate species.” In: *Science (New York, N.Y.)* (2012).

[25] Henry L Levin and John V Moran. “Dynamic interactions between transposable elements and their hosts.” In: *Nature reviews. Genetics* (2011).

[26] Christine R Beck et al. “LINE-1 elements in structural variation and disease.” In: *Annual review of genomics and human genetics* (2011).

[27] Nemanja Rodić and Kathleen H Burns. “Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms?” In: *PLoS genetics* (2013).

[28] Asaf Levy, Schraga Schwartz, and Gil Ast. “Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements”. In: *Nucleic acids research* (2009).

[29] I Ovchinnikov, A B Troxel, and G D Swergold. “Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion.” In: *Genome research* (2001).

[30] A Pavlícek et al. “Similar integration but different stability of Alus and LINEs in the human genome.” In: *Gene* (2001).

[31] G D Hurst and J H Werren. “The role of selfish genetic elements in eukaryotic evolution.” In: *Nature reviews. Genetics* (2001).

[32] D A Kramerov and N S Vassetzky. “Origin and evolution of SINEs in eukaryotic genomes.” In: *Heredity* (2011).

[33] Maria A Nilsson et al. “Tracking marsupial evolution using archaic genomic retroposon insertions.” In: *PLoS biology* (2010).

[34] Maria A Nilsson et al. “Expansion of CORE-SINEs in the genome of the Tasmanian devil.” In: *BMC genomics* (2012).

[35] Nikita S Vassetzky and Dmitri A Kramerov. “SINEBase: a database and tool for SINE analysis.” In: *Nucleic acids research* (2013).

[36] J Kim et al. “Rodent BC1 RNA gene as a master gene for ID element amplification.” In: *Proceedings of the National Academy of Sciences of the United States of America* (1994).

[37] J Kim and P L Deininger. “Recent amplification of rat ID sequences.” In: *Journal of molecular biology* (1996).

[38] Rhesus Macaque Genome Sequencing and Analysis Consortium et al. “Evolutionary and biomedical insights from the rhesus macaque genome.” In: *Science (New York, N.Y.)* (2007).

[39] Kyudong Han et al. “Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome.” In: *Science (New York, N.Y.)* (2007).

[40] E S Lander et al. “Initial sequencing and analysis of the human genome.” In: *Nature* (2001).

[41] Judith E Stenger et al. “Biased Distribution of Inverted and Direct Alus in the Human Genome: Implications for Insertion, Exclusion, and Genome Stability”. In: *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada* (2001).

[42] Joti Giordano et al. “Evolutionary history of mammalian transposons determined by genome-wide defragmentation.” In: *PLoS computational biology* (2007).

[43] Ying Zhang, Mark T Romanish, and Dixie L Mager. "Distributions of Transposable Elements Reveal Hazardous Zones in Mammalian Introns". In: *PLoS computational biology* (2011).

[44] Rotem Sorek, Gil Ast, and Dan Graur. "Alu-containing exons are alternatively spliced." In: *Genome research* (2002).

[45] G A Mitchell et al. "Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation." In: *Proceedings of the National Academy of Sciences of the United States of America* (1991).

[46] Galit Lev-Maor et al. "Intronic Alus influence alternative splicing." In: *PLoS genetics* (2008).

[47] Alekos Athanasiadis, Alexander Rich, and Stefan Maas. "Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome." In: *PLoS biology* (2004).

[48] Joost H A Martens et al. "The profile of repeat-associated histone lysine methylation states in the mouse epigenome." In: *The EMBO journal* (2005).

[49] Yutaka Kondo and Jean-Pierre J Issa. "Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells." In: *The Journal of biological chemistry* (2003).

[50] A C Arnberg, GJB Van Ommen, and L A Grivell. "Some yeast mitochondrial RNAs are circular". In: *Cell* (1980).

[51] M Danan et al. "Transcriptome-wide discovery of circular RNAs in Archaea". In: *Nucleic acids research* (2012).

[52] T O Diener. "Circular RNAs: relics of precellular evolution?" In: *Proceedings of the National Academy of ...* (1989).

[53] Marianne C Kramer et al. "Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and SR proteins." In: *Genes & development* (2015).

[54] J U Guo et al. "Expanded identification and characterization of mammalian circular RNAs". In: *Genome biology* (2014).

[55] Yehoshua Enuka et al. "Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor." In: *Nucleic acids research* (2016).

[56] Stefan Starke et al. "Exon circularization requires canonical splice signals." In: *Cell reports* (2015).

[57] R Ashwal-Fluss et al. "circRNA Biogenesis Competes with Pre-mRNA Splicing". In: *Molecular cell* (2014).

[58] D E Coulter and A L Greenleaf. "A mutation in the largest subunit of RNA polymerase II alters RNA chain elongation in vitro." In: *The Journal of biological chemistry* (1985).

[59] Ling-Ling Chen. "The biogenesis and emerging roles of circular RNAs." In: *Nature reviews. Molecular cell biology* (2016).

[60] R A Dubin, M A Kazmi, and H Ostrer. "Inverted repeats are necessary for circularization of the mouse testis Sry transcript." In: *Gene* (1995).

[61] Xiao-Ou Zhang et al. "Complementary sequence-mediated exon circularization." In: *Cell* (2014).

[62] Dongming Liang and Jeremy E Wilusz. "Short intronic repeat sequences facilitate circular RNA production." In: *Genes & development* (2014).

[63] Jeremy E Wilusz. "Repetitive elements regulate circular RNA biogenesis." In: *Mobile genetic elements* (2015).

[64] Andranik Ivanov et al. "Analysis of Intron Sequences Reveals Hallmarks of Circular RNA Biogenesis in Animals". In: *Cell reports* (2015).

[65] William R Jeck et al. "Circular RNAs are abundant, conserved, and associated with ALU repeats." In: *RNA (New York, N.Y.)* (2013).

[66] Agnieszka Rybak-Wolf et al. "Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed." In: *Molecular cell* (2015).

[67] Morten T Venø et al. "Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development." In: *Genome biology* (2015).

[68] Jakub O Westholm et al. "Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation." In: *Cell reports* (2014).

[69] Julia Salzman et al. "Cell-type specific features of circular RNA expression." In: *PLoS genetics* (2013).

[70] Yuan Gao et al. "Comprehensive identification of internal structure and alternative splicing events in circular RNAs." In: *Nature communications* (2016).

[71] Xiao-Ou Zhang et al. "Diverse alternative back-splicing and alternative splicing landscape of circular RNAs." In: *Genome research* (2016).

[72] Thomas B Hansen et al. "Natural RNA circles function as efficient microRNA sponges". In: *Nature* (2013).

[73] L F Thomas and P Saetrom. "Circular RNAs are depleted of polymorphisms at microRNA binding sites". In: *Bioinformatics (Oxford, England)* (2014).

[74] Giuseppe Militello et al. "Screening and validation of lncRNAs and circRNAs as miRNA sponges." In: *Briefings in Bioinformatics* (2016).

[75] Simon J Conn et al. "The RNA Binding Protein Quaking Regulates Formation of circRNAs." In: *Cell* (2015).

[76] William W Du et al. "Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2." In: *Nucleic acids research* (2016).

[77] Naoko Abe et al. "Rolling Circle Translation of Circular RNA in Living Human Cells." In: *Scientific reports* (2015).

[78] Yang Wang and Zefeng Wang. "Efficient backsplicing produces translatable circular mRNAs." In: *RNA (New York, N.Y.)* (2015).

[79] Yun Yang et al. "Extensive translation of circular RNAs driven by N(6)-methyladenosine." In: *Cell Research* (2017).

[80] Nagarjuna Reddy Pamudurti et al. "Translation of CircRNAs." In: *Molecular cell* (2017).

[81] Ivano Legnini et al. "Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis." In: *Molecular cell* (2017).

[82] Anna Bachmayr-Heyda et al. "Correlation of circular RNA abundance with proliferation–exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues." In: *Scientific reports* (2015).

[83] Haimin Li et al. "Circular RNA Expression Profile of Pancreatic Ductal Adenocarcinoma Revealed by Microarray." In: *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology* (2016).

[84] W J Lukiw. "Circular RNA (circRNA) in Alzheimer's disease (AD)". In: *Frontiers in genetics* (2013).

[85] Christin E Burd et al. "Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk." In: *PLoS genetics* (2010).

[86] Qiang Liu et al. "Circular RNA Related to the Chondrocyte ECM Regulates MMP13 Expression by Functioning as a MiR-136 'Sponge' in Human Cartilage Degradation." In: *Scientific reports* (2016).

[87] S Ghosal et al. "Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits". In: *Frontiers in genetics* (2013).

[88] Dawood B Dudekula et al. "CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs." In: *RNA biology* (2016).

[89] Anindya Bhattacharya and Yan Cui. "SomamiR 2.0: a database of cancer somatic mutations altering microRNA-ceRNA interactions." In: *Nucleic acids research* (2016).

[90] Sebastian Memczak et al. "Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood." In: *PloS one* (2015).

[91] J H Bahn et al. "The Landscape of MicroRNA, Piwi-Interacting RNA, and Circular RNA in Human Saliva". In: *Clinical Chemistry* (2014).

[92] Thomas B Hansen et al. "Comparison of circular RNA prediction tools." In: *Nucleic acids research* (2015).

[93] Sebastian Memczak et al. "Circular RNAs are a large class of animal RNAs with regulatory potency". In: *Nature* (2013).

[94] Yuan Gao, Jinfeng Wang, and Fangqing Zhao. "CIRI: an efficient and unbiased algorithm for de novo circular RNA identification." In: *Genome biology* (2015).

[95] Kai Wang et al. "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery." In: *Nucleic acids research* (2010).

[96] Yuan Gao, Jinyang Zhang, and Fangqing Zhao. "Circular RNA identification based on multiple seed matching." In: *Briefings in Bioinformatics* (2017).

[97] Linda Szabo and Julia Salzman. "Detecting circular RNAs: bioinformatic and experimental challenges." In: *Nature reviews. Genetics* (2016).

[98] Franziska Metge et al. "FUCHS-towards full circular RNA characterization using RNAseq." In: *PeerJ* (2017).

[99] Alexander Kanitz et al. "Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data." In: *Genome biology* (2015).

[100] Musheng Li et al. "Quantifying circular RNA expression from RNA-seq data using model-based framework." In: *Bioinformatics (Oxford, England)* (2017).

[101] Linda J Mullins and John J Mullins. "Insights from the rat genome sequence". In: *Genome biology* (2004).

[102] Anamaria Necsulea et al. "The evolution of lncRNA repertoires and expression patterns in tetrapods." In: *Nature* (2014).

[103] Hitoshi Suzuki et al. "Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing." In: *Nucleic acids research* (2006).

[104] C E Shannon. *The mathematical theory of communication. 1963.* 1997.

[105] Octavio Martínez and M Humberto Reyes-Valdés. "Defining diversity, specialization, and gene specificity in transcriptomes through information theory." In: *Proceedings of the National Academy of Sciences* (2008).

[106] Adam Siepel et al. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." In: *Genome research* (2005).

[107] Atsushi Ogura, Kazuho Ikeo, and Takashi Gojobori. "Comparative analysis of gene expression for convergent evolution of camera eye between octopus and human." In: *Genome research* (2004).

[108] Meredith E Protas and Nipam H Patel. "Evolution of Coloration Patterns". In: *Annual Review of Cell and Developmental Biology* (2008).

[109] Liucun Zhu et al. "Patterns of exon-intron architecture variation of genes in eukaryotic genomes." In: *BMC genomics* (2009).

[110] Z Zhang et al. "A greedy algorithm for aligning DNA sequences." In: *Journal of computational biology : a journal of computational molecular cell biology* (2000).

[111] Amnon Koren et al. "Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation". In: *The American Journal of Human Genetics* (2012).

[112] Athma A Pai et al. "The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels." In: *PLoS genetics* (2012).

[113] Julia Steinberg et al. "Haploinsufficiency predictions without study bias." In: *Nucleic acids research* (2015).

[114] Chun-Long Chen et al. "Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes." In: *Genome research* (2010).

[115] John A Stamatoyannopoulos et al. "Human mutation rate associated with DNA replication timing." In: *Nature genetics* (2009).

[116] M Richard Shen, Mark A Batzer, and Prescott L Deininger. "Evolution of the master Alu gene(s)". In: *Journal of Molecular Evolution* (1991).

[117] Vladimir V Kapitonov and Jerzy Jurka. "A universal classification of eukaryotic transposable elements implemented in Repbase." In: *Nature reviews. Genetics* (2008).

[118] Ronny Lorenz et al. "ViennaRNA Package 2.0." In: *Algorithms for molecular biology : AMB* (2011).

[119] Gero Doose et al. "Mapping the RNA-Seq trash bin: unusual transcripts in prokaryotic transcriptome sequencing data." In: *RNA biology* (2013).

[120] Karoline K Ebbesen, Jørgen Kjems, and Thomas B Hansen. "Circular RNAs: Identification, biogenesis and function." In: *Biochimica et biophysica acta* (2015).

[121] Y Zhang et al. "Circular Intronic Long Noncoding RNAs". In: *Molecular cell* (2013).

[122] Petar Glažar, Panagiotis Papavasileiou, and Nikolaus Rajewsky. "circBase: a database for circular RNAs". In: *RNA (New York, N.Y.)* (2014).

[123] Xintian You et al. "Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity." In: *Nature neuroscience* (2015).

[124] Peter L Wang et al. "Circular RNA is expressed across the eukaryotic tree of life." In: *PloS one* (2014).

[125] Rumana Bahar et al. “Increased cell-to-cell variation in gene expression in ageing mouse heart.” In: *Nature* (2006).

[126] Michael A Lodato et al. “Somatic mutation in single human neurons tracks developmental and transcriptional history.” In: *Science (New York, N.Y.)* (2015).

[127] D G de Rooij and L D Russell. “All you wanted to know about spermatogonia but were afraid to ask.” In: *Journal of andrology* (2000).

[128] R Lamprecht and J LeDoux. “Structural plasticity and memory : Abstract : Nature Reviews Neuroscience”. In: *Nature Reviews Neuroscience* (2004).

[129] J A Yoder, C P Walsh, and T H Bestor. “Cytosine methylation and the ecology of intragenomic parasites.” In: *Trends in genetics : TIG* (1997).

[130] Tuğçe Aktaş et al. “DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome.” In: *Nature* (2017).

[131] Dmitrii Polev. “Transcriptional noise as a driver of gene evolution.” In: *Journal of theoretical biology* (2012).

[132] J Brosius and S J Gould. “On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA"”. In: *... of the National Academy of Sciences* (1992).

[133] Florent Campo-Paysaa et al. “microRNA complements in deuterostomes: origin and evolution of microRNAs.” In: *Evolution & development* (2011).

[134] Lucía F Franchini et al. “Convergent evolution of two mammalian neuronal enhancers by sequential exaptation of unrelated retroposons.” In: *Proceedings of the National Academy of Sciences* (2011).

[135] Cole Trapnell et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.” In: *Nature protocols* (2012).

[136] A J Enright, S Van Dongen, and C A Ouzounis. “An efficient algorithm for large-scale detection of protein families.” In: *Nucleic acids research* (2002).

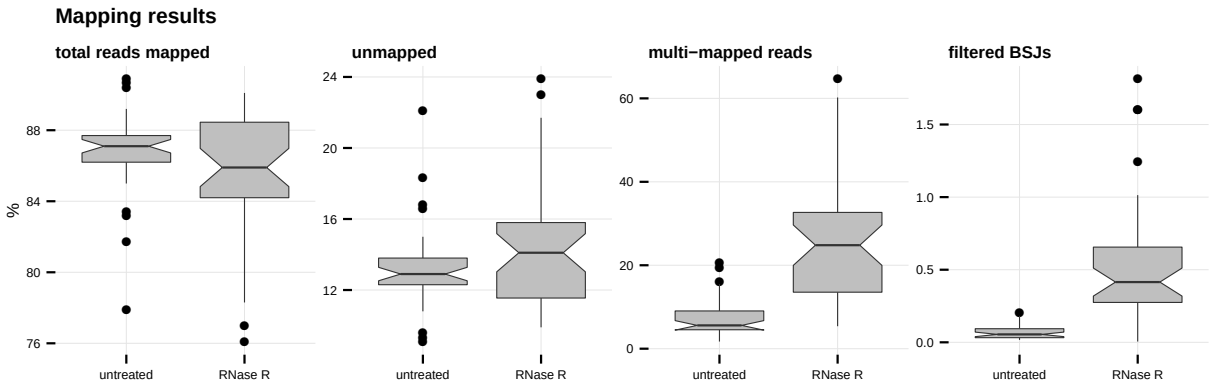
[137] Stijn Dongen. “Performance criteria for graph clustering and Markov cluster experiments”. In: (2000).

[138] Eran Eden et al. “GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.” In: *BMC bioinformatics* (2009).

6 Supplementary Data

Supplementary Table 1: Sample overview | For each sample, information on the organism, internal ID, tissue, age and sex are summarized. Last row shows the RNA Quality Number (RQN) for the extracted RNA.

Species	Internal ID	Tissue	Age	Sex	RQN
Opossum	OPO 1270 cerebellum	Cerebellum	21 months	male	7.3
Opossum	OPO 1272 cerebellum	Cerebellum	19.5 months	male	8.9
Opossum	OPO 1359 cerebellum	Cerebellum	15.5 months	male	6.8
Opossum	OPO 1359 liver	Liver	15.5 months	male	9.3
Opossum	OPO 1270 liver	Liver	21 months	male	8.6
Opossum	OPO 1298 liver	Liver	13 months	male	9
Opossum	OPO 1270 testis	Testis	21 months	male	8.9
Opossum	OPO 1298 testis	Testis	13 months	male	8.5
Opossum	OPO 1359 testis	Testis	15.5 months	male	8.9
Mouse	MOU 9W M1 020713 Cr	Cerebellum	9 weeks	male	7.1
Mouse	MOU 9W M3 020713 Cr	Cerebellum	9 weeks	male	7.4
Mouse	MOU 9W M4 020713 Cr	Cerebellum	9 weeks	male	7
Mouse	MOU 9W M1 020713 Lv	Liver	9 weeks	male	7.9
Mouse	MOU 9W M3 020713 Lv	Liver	9 weeks	male	7.6
Mouse	MOU 9W M4 020713 Lv	Liver	9 weeks	male	8.5
Mouse	MOU 9W M1 020713 GT	Testis	9 weeks	male	8.4
Mouse	MOU 9W M3 020713 GT	Testis	9 weeks	male	8.2
Mouse	MOU 9W M4 020713 GT	Testis	9 weeks	male	8.4
Rat	RAT 16W M3 161013 Cr	Cerebellum	16 weeks	male	7.2
Rat	RAT 16W M2 161013 Cr	Cerebellum	17 weeks	male	7.5
Rat	RAT 16W M1 161013 Cr	Cerebellum	18 weeks	male	7.7
Rat	RAT 16W M3 161013 Lv	Liver	19 weeks	male	7.2
Rat	RAT 16W M2 161013 Lv	Liver	20 weeks	male	7.9
Rat	RAT 16W M1 161013 Lv	Liver	21 weeks	male	7.8
Rat	RAT 16W M3 161013 GT	Testis	22 weeks	male	7.7
Rat	RAT 16W M2 161013 GT	Testis	23 weeks	male	8.8
Rat	RAT 16W M1 161013 GT	Testis	24 weeks	male	7.8
Rhesus	MAC 99057 M1 180414 Br	Cerebellum	8 years	male	8.5
Rhesus	MAC 98073 M3 180414 Br	Cerebellum	9 years	male	7.7
Rhesus	MAC 99057 M1 180414 Lv	Liver	8 years	male	8.6
Rhesus	MAC 98145 M2 180414 Lv	Liver	9 years	male	8.2
Rhesus	MAC 98073 M3 180414 Lv	Liver	9 years	male	8.6
Rhesus	MAC 99057 M1 180414 GT	Testis	8 years	male	9.5
Rhesus	MAC 98145 M2 180414 GT	Testis	9 years	male	9.1
Rhesus	MAC 99057 M1 180414 GT	Testis	8 years	male	8.8
Human	virtual HUM 361 M	Liver	64 years	male	7.5
Human	HUM 1502 M4 180414 Br	Cerebellum	29 years	male	8.2
Human	HUM 1134 M6 180414 Br	Cerebellum	41 years	male	8.6
Human	HUM 605 M8 180414 Br	Cerebellum	25 years	male	8.3
Human	HUM 1403 M5 180414 GT	Testis	21 years	male	7.8
Human	HUM 1134 M6 180414 GT	Testis	41 years	male	6.9
Human	HUM 1027 M7 180414 GT	Testis	22 years	male	6.9



Supplementary Figure 1: Mapping summary | Plots show the percentage of mapped, unmapped, multi-mapped and BSJ reads across all libraries in untreated and RNase R treated conditions.

Supplementary Table 2: Detected BSJs across samples | Table summarizes the total number of detected BSJs after filtering in each species. The percentage of BSJs that are unique to one, two, three or more than three samples is shown.

Species	Total BSJs	1 replicate	2 replicates	3 replicates	>= 4 replicates
Opossum	76,739	84.74	8.05	4.28	2.93
Mouse	67,249	83.45	9.23	4.73	2.59
Rat	72,855	85.43	7.73	3.88	2.96
Rhesus macaque	100,270	79.29	9.79	4.83	6.09
Human	68,400	79.86	10.71	6.54	2.9

Supplementary Table 3: Total number of circRNAs in different species and tissues | Indicated is the total number of different circRNAs that were annotated in each of the tissues across all species.

Species	Liver	Cerebellum	Testis
Opossum	129	417	1229
Mouse	87	1054	523
Rat	114	996	1192
Rhesus macaque	601	2132	1367
Human	765	2994	1761

Supplementary Table 4: CircRNAs confirmed by qPCR | CircRNAs validated by qPCR for different species and tissues. Candidates were selected based on high expression in a given tissue.

Gene name	Species	Tissue	Validated
Adk	Mouse	Liver	Yes
Adk	Mouse	Liver	Yes
Ankib1	Mouse	Liver	Yes
Apoa2	Mouse	Liver	No
Arhgap5	Mouse	Liver	No
Asph	Mouse	Liver	Yes
Cdyl	Mouse	Cerebellum	Yes
Dtnb	Mouse	Liver	Yes
Dnah7a	Mouse	Testis	Yes
Evi5	Mouse	Liver	Yes
Gigyf2	Mouse	Liver	Yes
Gm21992/Rbm4	Mouse	Liver	Yes
Homer1	Mouse	Liver	Yes
Med13l	Mouse	Cerebellum	Yes
Nfix	Mouse	Liver	Yes
Nr1h4	Mouse	Liver	Yes
Prox1	Mouse	Liver	Yes
Rabep1	Mouse	Liver	Yes
Rad52	Mouse	Liver	Yes
Rbm33	Mouse	Liver	Yes
Rere	Mouse	Cerebellum	Yes
Rmdn2	Mouse	Liver	Yes
Rnf169	Mouse	Liver	Yes
Scaper	Mouse	Liver	Yes
Strn3	Mouse	Liver	Yes
Tasp1	Mouse	Testis	Yes
Tmem56	Mouse	Liver	Yes
Ube2k	Mouse	Liver	Yes
Rere	Rat	Cerebellum	Yes
Rere	Human	Cerebellum	Yes

Supplementary Table 5: Median GC content of different exon types | Grand median GC content for exons inside and outside the circRNA for each isochore and species (1 corresponding to 100%). The difference between circRNA-contained and circRNA-outside exons was assessed with a one-tailed and paired Mann-Whitney U test, p-values are indicated. *Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

Species	Isochore	GC in circRNA-contained exons	GC in circRNA-outside exons	p-value
Opossum	L1	0.41	0.42	0.00063
Opossum	L2	0.44	0.46	0.00003
Opossum	H1	0.50	0.51	0.02331
Opossum	H2	0.51	0.60	0.00977
Opossum	H3	0.57	0.64	0.25000
Mouse	L1	0.43	0.43	0.38241
Mouse	L2	0.44	0.46	0.00000
Mouse	H1	0.49	0.51	0.00002
Mouse	H2	0.54	0.56	0.00206
Mouse	H3	0.56	0.59	0.01636
Rat	L1	0.41	0.42	0.13878
Rat	L2	0.44	0.46	0.00000
Rat	H1	0.49	0.50	0.02392
Rat	H2	0.54	0.55	0.00658
Rat	H3	0.59	0.58	0.59608
Rhesus macaque	L1	0.41	0.41	0.55077
Rhesus macaque	L2	0.43	0.44	0.00002
Rhesus macaque	H1	0.47	0.50	0.00000
Rhesus macaque	H2	0.54	0.56	0.01153
Rhesus macaque	H3	0.58	0.61	0.00011
Human	L1	0.41	0.42	0.01364
Human	L2	0.43	0.44	0.00000
Human	H1	0.47	0.50	0.00000
Human	H2	0.54	0.56	0.00000
Human	H3	0.58	0.61	0.00008

Supplementary Table 6: Mean amplitude correlations | Spearman's rank correlation for the GC amplitude and GC content of introns and exons were calculated for each isochore and species. The mean correlation between the GC amplitude and GC content of introns and exons for different splice sites relative to the circRNA is shown.

Position	Amplitude ~ Intron	Amplitude ~ Exon
Non-parental	-0.42	0.31
Outside circRNA	-0.44	0.16
Inside circRNA	-0.48	0.40

Supplementary Table 7: GLM summary for presence of a parental gene | A GLM was fitted on all coding genes incorporating the different predictors described in Table 6 ($n_{opossum}$=18,807, n_{mouse}=22,015, n_{rat}=11,654, n_{rhesus}=21,891, n_{human}=21,744). Table provides information on the effect size of each predictor, its confidence intervals (lower and upper CI) and the significance. *Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

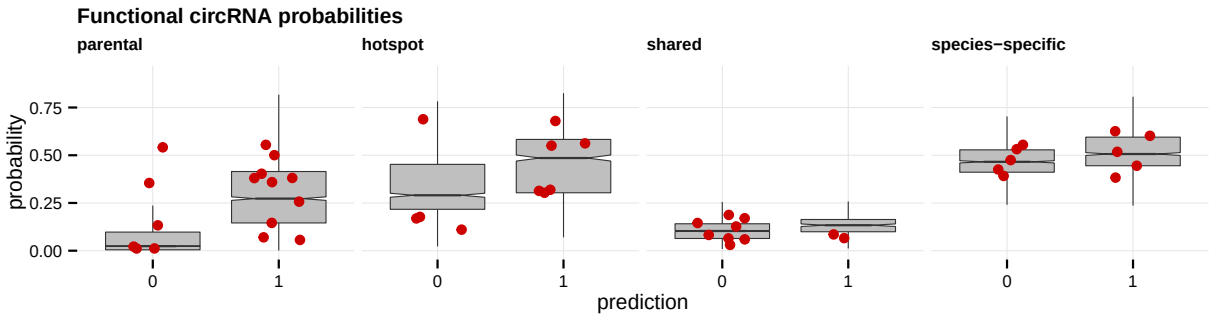
Species	Predictor	Odds ratio	Upper CI	Lower CI	p-value
opossum	as.rvc	0.2124	0.1395	0.2868	0.0
opossum	exon_count	0.31	0.2587	0.3613	0.0
opossum	genomic_length	0.2554	0.1485	0.367	0.0
opossum	median_brawand	0.0808	0.0107	0.1215	0.0001
opossum	percentage_gc_content	-1.8545	-2.0544	-1.6622	0.0
opossum	percentage_gc_content:genomic_length	0.2063	0.0908	0.3277	0.0006
opossum	ss.rvc	0.1109	0.0171	0.1949	0.0143
mouse	as.repeats/genomic_length	0.3553	0.2535	0.4553	0.0
mouse	exon_count	0.3087	0.2452	0.3714	0.0
mouse	genomic_length	0.4852	0.4008	0.5698	0.0
mouse	percentage_gc_content	-0.9959	-1.111	-0.8834	0.0
mouse	percentage_gc_content:genomic_length	0.3001	0.2301	0.3712	0.0
mouse	phastcons	0.8546	0.7407	0.9731	0.0
mouse	ss.repeats/genomic_length	0.1136	0.0232	0.203	0.0132
mouse	transcript_count	0.3388	0.2538	0.4246	0.0
mouse	transcript_count:exon_count	-0.087	-0.118	-0.059	0.0
rat	as.rvc	0.2626	0.1493	0.3752	0.0
rat	as.rvc:ss.rvc	-0.0673	-0.109	-0.033	0.0006
rat	exon_count	0.1605	0.1025	0.2179	0.0
rat	genomic_length	0.4332	0.3558	0.5105	0.0
rat	mean_cpm	0.0682	0.016	0.1236	0.0082
rat	percentage_gc_content	-0.7951	-0.9209	-0.6717	0.0
rat	percentage_gc_content:genomic_length	0.2868	0.2157	0.36	0.0
rat	phastcons	0.6542	0.5132	0.8033	0.0
rat	ss.rvc	0.1735	0.0783	0.262	0.0002
rat	transcript_count	0.0668	0.0174	0.1145	0.0069
rhesus	as.rvc	0.6424	0.5888	0.6962	0.0
rhesus	exon_count	0.3556	0.3039	0.4074	0.0
rhesus	genomic_length	0.3367	0.2539	0.4192	0.0
rhesus	percentage_gc_content	-1.2782	-1.3788	-1.1802	0.0
rhesus	percentage_gc_content:genomic_length	0.1672	0.0912	0.2427	0.0
rhesus	transcript_count	0.055	0.0058	0.1035	0.0273
rhesus	tsi	-0.0649	-0.1249	-0.005	0.034
human	as.rvc	0.3624	0.2449	0.4838	0.0
human	exon_count	0.2917	0.233	0.3505	0.0
human	genomic_length	0.4297	0.3469	0.513	0.0
human	percentage_gc_content	-1.2136	-1.3074	-1.1221	0.0
human	percentage_gc_content:genomic_length	0.2981	0.2287	0.3681	0.0
human	phastcons	0.6003	0.5327	0.6689	0.0
human	ss.rvc	0.2877	0.1637	0.4066	0.0
human	transcript_count	0.3491	0.2804	0.4174	0.0
human	transcript_count:exon_count	-0.0898	-0.1154	-0.0653	0.0

Supplementary Table 8: GLM summary for presence of a parental hotspot gene | A GLM was fitted on all parental hotspot genes incorporating the different predictors described in Table 6 ($n_{opossum}$=884, n_{mouse}=858, n_{rat}=983, n_{rhesus}=1704, n_{human}=2058). Table provides information on the effect size of each predictor, its confidence intervals (lower and upper CI) and the significance. *Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

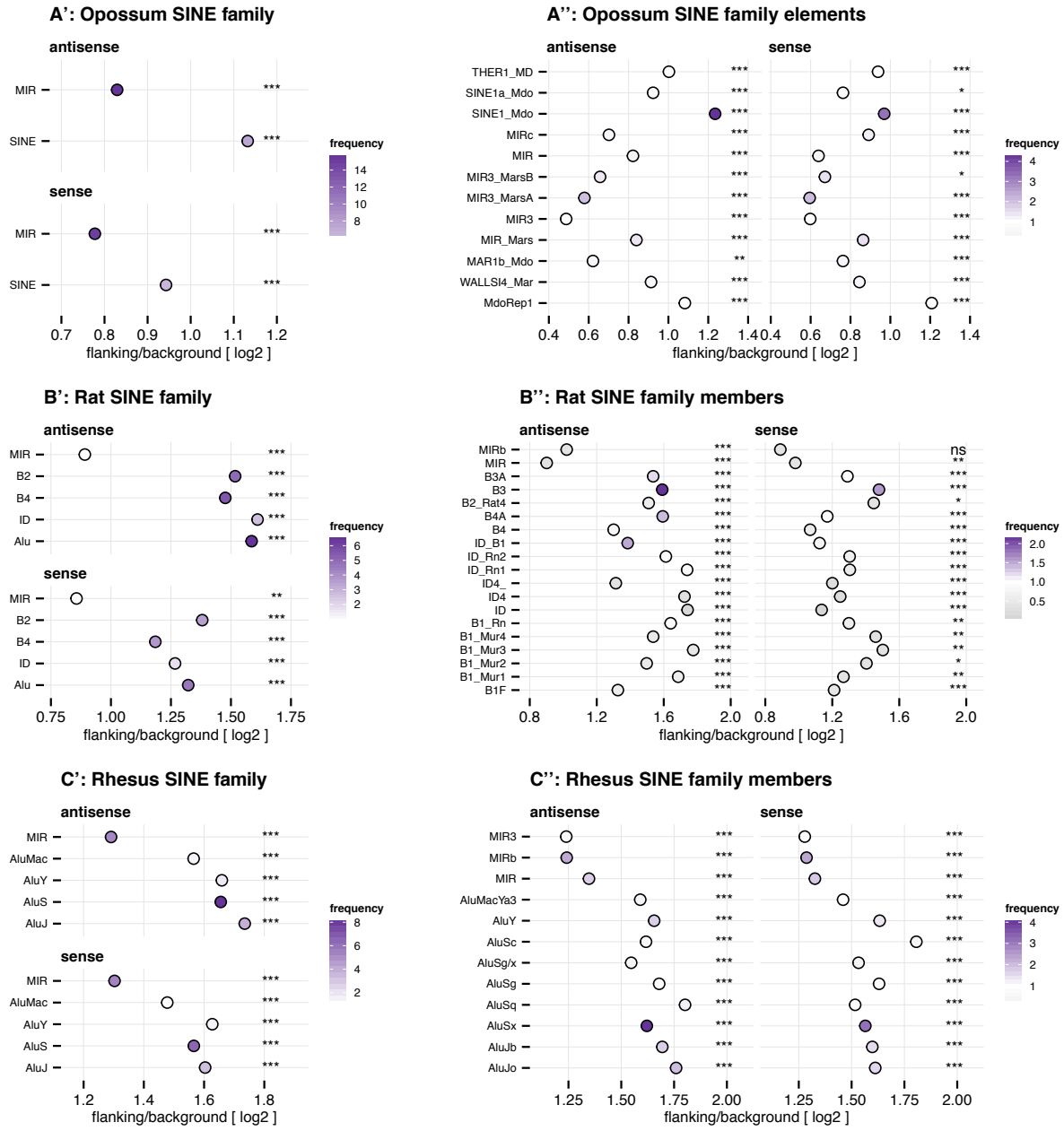
Species	Predictor	Odds ratio	Upper CI	Lower CI	p-value
opossum	ageopossum	-0.8453	-1.2057	-0.4839	0.0
opossum	percentage_gc_content	-0.5566	-0.8418	-0.2969	0.0001
opossum	tsi	0.1909	0.0188	0.3622	0.029
mouse	ageeutherian	-0.5059	-0.9408	-0.0716	0.0224
mouse	agemouse	-1.9525	-2.4477	-1.4702	0.0
mouse	agerodents	-1.0478	-1.7001	-0.4274	0.0012
mouse	as.rvc	-0.6425	-0.9873	-0.3201	0.0001
mouse	phastcons	0.2301	0.0326	0.4426	0.0274
mouse	ss.rvc	0.8193	0.4986	1.1688	0.0
mouse	transcript_count	0.2543	0.0956	0.4163	0.0018
rat	ageeutherian	-0.8431	-1.2831	-0.4092	0.0002
rat	agerat	-1.9521	-2.3869	-1.5258	0.0
rat	agerodents	-1.3726	-2.0288	-0.7489	0.0
rat	percentage_gc_content	-0.2806	-0.4602	-0.1077	0.0018
rat	ss.rvc	0.1937	0.0474	0.3395	0.009
rhesus	ageeutherian	0.0639	-0.2902	0.4182	0.7235
rhesus	ageprimates	-0.485	-0.8093	-0.1615	0.0033
rhesus	agerhesus	-1.4125	-1.742	-1.0854	0.0
rhesus	percentage_gc_content	-0.1872	-0.306	-0.0717	0.0017
human	ageeutherian	-0.6614	-1.0267	-0.3015	0.0003
human	agehuman	-1.806	-2.131	-1.489	0.0
human	ageprimates	-0.8923	-1.2341	-0.5572	0.0
human	median_brawand	0.1075	0.004	0.2093	0.039
human	percentage_gc_content	-0.486	-0.6125	-0.364	0.0
human	ss.rvc	0.2348	0.1346	0.3357	0.0

Supplementary Table 9: GLM summary for shared and species-specific circRNA loci | A: Presence of a shared circRNA locus. A GLM was fitted on all parental genes incorporating the different predictors described in Table 6 ($n_{opossum}$=884, n_{mouse}=858, n_{rat}=983, n_{rhesus}=1704, n_{human}=2058). Table provides information on the effect size of each predictor, its confidence intervals (lower and upper CI) and the significance. **B:** Presence of a species-specific circRNA locus. *Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

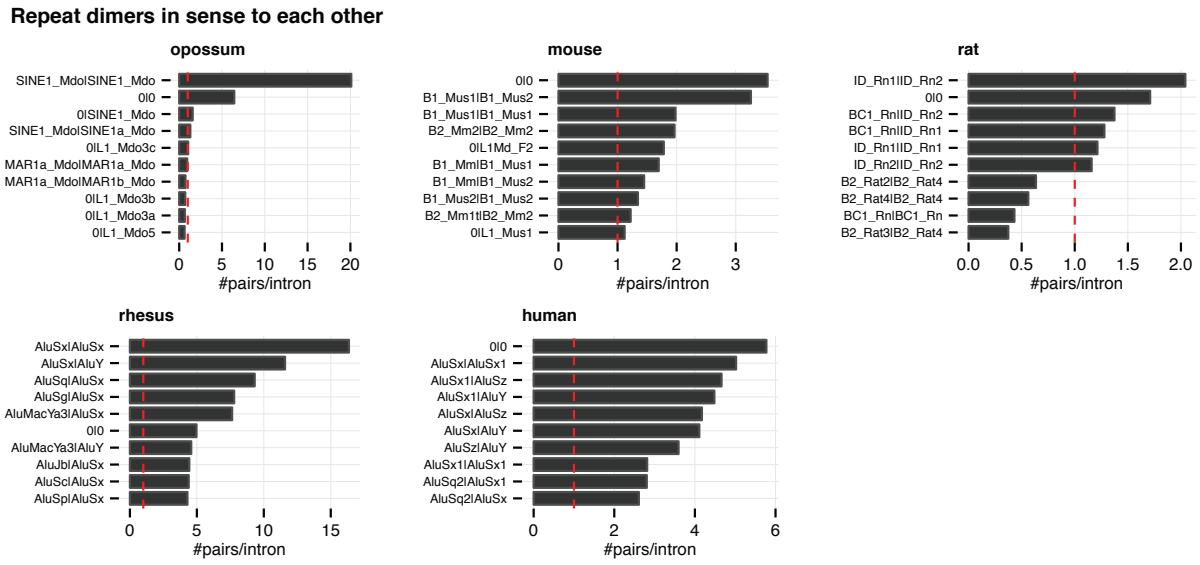
A:	Species	Predictor	Odds ratio	Upper CI	Lower CI	p-value
	opossum	genomic_length	0.2851	0.1393	0.434	0.0001
	opossum	percentage_gc_content	-0.5597	-0.801	-0.3368	0.0
	opossum	tsi	-0.3208	-0.4878	-0.1589	0.0001
	mouse	percentage_gc_content	-0.3892	-0.6087	-0.1822	0.0003
	mouse	transcript_count	0.2244	0.0585	0.3883	0.0071
	rat	genomic_length	0.175	0.0086	0.3328	0.0333
	rat	percentage_gc_content	-0.3132	-0.5354	-0.1023	0.0045
	rat	phastcons	0.2887	0.0494	0.5759	0.0301
	rhesus	genomic_length	0.2367	0.0867	0.3821	0.0014
	rhesus	percentage_gc_content	-0.355	-0.5426	-0.1776	0.0001
	rhesus	percentage_gc_content:genomic_length	0.1546	-0.0004	0.3133	0.0473
	rhesus	transcript_count	0.1353	-0.0007	0.2657	0.046
	human	percentage_gc_content	-0.427	-0.6156	-0.2494	0.0
	human	phastcons	0.3763	0.2041	0.5613	0.0
	human	transcript_count	0.2147	0.0924	0.3348	0.0005
B:	Species	Predictor	Odds ratio	Upper CI	Lower CI	p-value
	opossum	genomic_length	-0.4046	-0.5809	-0.2365	0.0
	opossum	percentage_gc_content	0.5716	0.3505	0.8106	0.0
	opossum	ss.rvc	0.3519	0.1247	0.5957	0.0034
	opossum	tsi	0.2953	0.1308	0.4648	0.0005
	mouse	exon_count	0.2134	0.057	0.3735	0.008
	mouse	percentage_gc_content	0.3023	0.1605	0.447	0.0
	mouse	phastcons	-0.3279	-0.4779	-0.1823	0.0
	mouse	transcript_count	-0.3117	-0.4854	-0.1471	0.0003
	rat	genomic_length	-0.2435	-0.3874	-0.1062	0.0007
	rat	percentage_gc_content	0.233	0.0993	0.3694	0.0007
	rat	phastcons	-0.3823	-0.541	-0.2344	0.0
	rhesus	as.rvc	-0.1462	-0.2484	-0.0449	0.0049
	rhesus	genomic_length	-0.1148	-0.2335	-0.0037	0.0495
	rhesus	percentage_gc_content	0.2832	0.1792	0.3891	0.0
	rhesus	transcript_count	-0.158	-0.2651	-0.0536	0.0034
	rhesus	tsi	0.1928	0.0935	0.2926	0.0001
	human	percentage_gc_content	0.2283	0.1368	0.3213	0.0
	human	phastcons	-0.2783	-0.3714	-0.1865	0.0
	human	transcript_count	-0.2492	-0.3463	-0.1549	0.0



Supplementary Figure 2: GLM probabilities of functional circRNAs | Functional circRNAs produced by human coding genes were selected from literature (WDR37, NFATC3, ITCH, APEX1, PVRL1, MYL12B, SIL1, GRSF1, PDPK1, PIK3AP1, ZNF524, RAB40A, PRNT, NR2C2AP, FAM110C, FAM78B). The probability of their parental gene to be parental, a hotspot, shared or species-specific is plotted in respect to the probability of all genes assessed (grey boxplot).



Supplementary Figure 3: Repeat frequency in flanking and background introns | A': Opossum SINE families. Plotted is the log2-enrichment for different repeat families in antisense and sense orientation. Increase in color intensity reflects the mean number of repeats detected in all flanking introns. Significance was estimated with a one-tailed Mann-Whitney U test. **A'':** Opossum SINE family elements. Plotted is the log2-enrichment for individual SINE family members in antisense and sense orientation. Increase in color intensity reflects the mean number of repeats detected in all flanking introns. Significance was estimated with a one-tailed Mann-Whitney U test. Family members that were not significant in sense and antisense orientation were removed from plot. **B':** Rat SINE families. **B'':** Rat SINE family elements. **C':** Rhesus SINE families. **C'':** Rhesus SINE family elements. *Abbreviations: as = antisense, ss = sense. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*



Supplementary Figure 4: Repeat dimers in sense to each other | Frequencies of the Top 10 repeat dimers found in flanking introns of each species. Red line was set at position 1 to indicate which repeats occur less or more often than once in the flanking introns. "0" in a dimer name corresponds to alignments for which no overlapping repeat was found after the intersection with the RepeatMasker annotation.

Supplementary Table 10: Minimal free energy for TE dimers | MFE (kcal/mol) of the secondary structure of the SINE TE dimers was calculated with the RNAcofold function from the ViennaRNA package. The reference TE was chosen based on its frequency in the dimer formation.

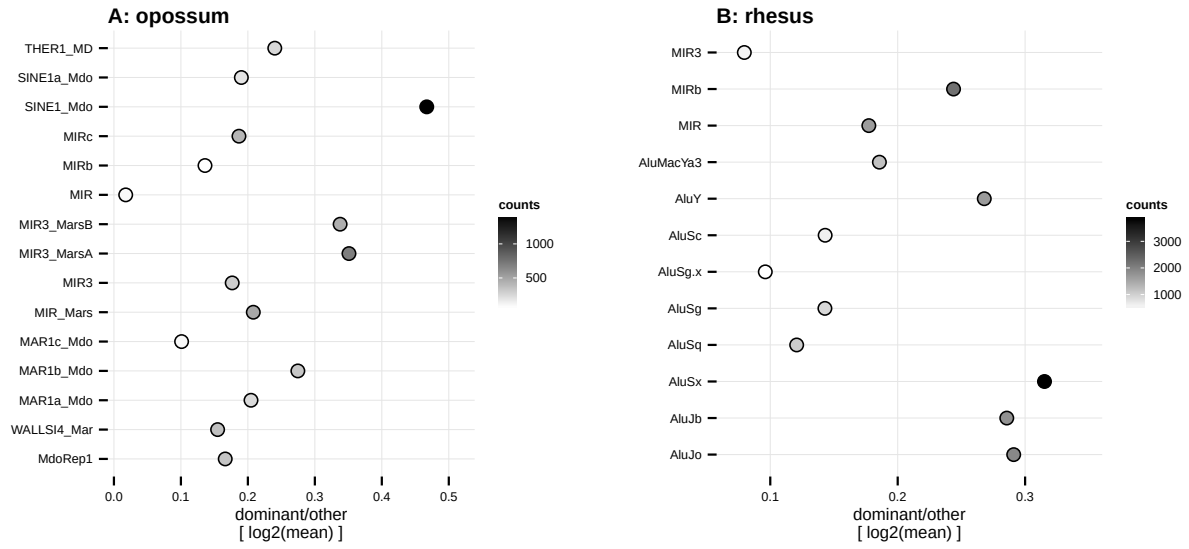
Species	Reference TE	Target TE	MFE
opossum	SINE1_Mdo	SINE1_Mdo	-453.30
opossum	SINE1_Mdo	MAR1_MD	-274.40
opossum	SINE1_Mdo	MAR1	-251.50
opossum	SINE1_Mdo	MIR3_MarsB	-230.20
opossum	SINE1_Mdo	MIR3_MarsA	-218.80
opossum	SINE1_Mdo	THER1_MD	-258.30
mouse	B1_Mus1	B1_Mus1	-333.20
mouse	B1_Mus1	B1_Mus2	-327.30
mouse	B1_Mus1	B1_Mm	-318.50
mouse	B1_Mus1	B1	-307.00
mouse	B1_Mus1	B1_Mur4	-320.80
mouse	B1_Mus1	B1_Mur3	-319.50
mouse	B1_Mus1	B1_Mur2	-297.80
mouse	B1_Mus1	B1_Mur1	-292.70
mouse	B1_Mus1	B1F2	-281.30
mouse	B1_Mus1	B1F1	-268.30
mouse	B1_Mus1	B1F	-245.00
mouse	B1_Mus1	B1-dID	-211.80
mouse	B1_Mus1	B4	-273.90
mouse	B1_Mus1	B4A	-264.50
mouse	B1_Mus1	ID_B1	-253.80
rat	ID_Rn1	ID_Rn1	-204.40
rat	ID_Rn1	ID_Rn2	-178.50
rat	ID_Rn1	BC1_Rn	-173.80
rat	ID_Rn1	BC1	-145.10
rat	ID_Rn1	ID_B1	-191.60
rhesus	AluSx	AluMacYb4	-567.20
rhesus	AluSx	AluMacYb2	-580.20
rhesus	AluSx	AluMacYa3	-587.90
rhesus	AluSx	AluY	-618.50
rhesus	AluSx	AluSc	-633.90
rhesus	AluSx	AluSg	-664.40
rhesus	AluSx	AluSq	-654.50
rhesus	AluSx	AluSx	-674.70
rhesus	AluSx	AluSz	-668.10
rhesus	AluSx	AluJb	-611.60
rhesus	AluSx	AluJo	-584.60
rhesus	AluSx	AluJr	-582.50
human	AluSx	AluY	-618.50
human	AluSx	AluSc	-633.90
human	AluSx	AluSg	-664.40
human	AluSx	AluSq2	-656.00
human	AluSx	AluSq	-654.50
human	AluSx	AluSx1	-668.40
human	AluSx	AluSx	-674.70
human	AluSx	AluSz	-668.10
human	AluSx	AluJb	-611.60
human	AluSx	AluJo	-584.60
human	AluSx	AluJr	-582.50

Supplementary Table 11: GLM summary for hotspot presence and depth | A: Presence of a hotspot. A GLM was fitted on all circRNA loci incorporating the different predictors described in Table 8 ($n_{opossum}$=1049, n_{mouse}=1024, n_{rat}=1285, n_{rhesus}=2169, n_{human}=2759). Table provides information on the effect size of each predictor, its confidence intervals (lower and upper CI) and the significance. **B:** Depth of hotspot. GLM analysis was restricted to hostpots with a maximum of five circRNAs ($n_{opossum}$=203, n_{mouse}=234, n_{rat}=305, n_{rhesus}=605, n_{human}=846). *Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

A:	Species	Predictor	Odds ratio	Lower CI	Upper CI	p-value
	opossum	gc	-0.849	-1.1469	-0.5713	0.0
	opossum	total.repeats	2.2674	1.9177	2.6452	0.0
	opossum	length	1.0351	0.8517	1.2273	0.0
	opossum	dG.500	0.2354	0.0147	0.482	0.0324
	mouse	gc	-0.1988	-0.4317	0.0282	0.0898
	mouse	total.repeats	1.8983	1.5735	2.2498	0.0
	mouse	length	1.2752	1.0703	1.4944	0.0
	mouse	age.rodents	1.527	0.8745	2.1745	0.0
	mouse	age.eutherian	1.0059	0.5274	1.4945	0.0
	mouse	age.therian	0.9069	0.3096	1.5017	0.0028
	rat	gc	-0.1928	-0.3875	-0.0022	0.0496
	rat	total.repeats	1.6646	1.4018	1.9462	0.0
	rat	length	1.2499	1.0628	1.4483	0.0
	rat	dG.500	0.16	0.011	0.2795	0.0098
	rat	age.rodents	1.0551	0.4447	1.6438	0.0005
	rat	age.eutherian	0.6386	0.2355	1.039	0.0018
	rat	age.therian	1.3411	0.854	1.8276	0.0
	rhesus	gc	-0.8067	-1.0056	-0.6166	0.0
	rhesus	total.repeats	3.0288	2.713	3.3635	0.0
	rhesus	length	1.0546	0.9108	1.2042	0.0
	rhesus	dG.250	0.1481	-0.0033	0.2508	0.0059
	rhesus	age.primates	0.6659	0.3446	0.9881	0.0
	rhesus	age.eutherian	0.7696	0.3819	1.1548	0.0001
	rhesus	age.therian	0.6726	0.2435	1.0966	0.002
	human	gc	-1.0955	-1.2861	-0.9134	0.0
	human	total.repeats	3.5266	3.2022	3.8693	0.0
	human	length	0.9933	0.8626	1.1287	0.0
	human	age.primates	0.6479	0.3616	0.9331	0.0
	human	age.eutherian	0.7718	0.4413	1.1003	0.0
	human	age.therian	0.9017	0.4929	1.3091	0.0
B:	Species	predictor	Odds ratio	Lower CI	Upper CI	p-value
	opossum	total.repeats	0.1166	0.037	0.1913	0.003
	opossum	length	0.1564	0.0719	0.2396	0.0003
	mouse	total.repeats	0.147	0.0726	0.2184	0.0001
	mouse	length	0.1249	0.0459	0.2021	0.0017
	rat	total.repeats	0.1178	0.0506	0.1827	0.0005
	rat	length	0.0963	0.0259	0.1654	0.0068
	rhesus	gc	-0.0904	-0.1509	-0.0316	0.003
	rhesus	total.repeats	0.1686	0.1229	0.2128	0.0
	rhesus	length	0.0997	0.0491	0.1497	0.0001
	human	gc	-0.0585	-0.1095	-0.0085	0.0232
	human	total.repeats	0.1746	0.1323	0.2162	0.0
	human	length	0.1111	0.0685	0.1532	0.0

Supplementary Table 12: LMM for circRNA dominance | A linear mixed model was fitted on all hotspots with a maximum of five circRNAs incorporating the different predictors described in Table 8 ($n_{opossum}$=203, n_{mouse}=234, n_{rat}=305, n_{rhesus}=605, n_{human}=846). Table provides information on the effect size of each predictor, its confidence intervals (lower and upper CI) and the significance.

Species	Predictor	Odds ratio	Lower CI	Upper CI	p-value
opossum	genomic.length	-0.0411	-0.0659	-0.0659	0.0012
mouse	genomic.length	-0.0305	-0.0534	-0.0075	0.0098
mouse	mean.dist	0.0279	0.0049	0.0508	0.0181
mouse	circ.age.rodents	0.0849	0.0191	0.1508	0.012
mouse	circ.age.eutherian	0.1136	0.0504	0.1768	0.0005
mouse	circ.age.therian	0.1169	-0.0039	0.2378	0.0593
rat	genomic.length	-0.0205	-0.0407	-0.0003	0.0479
rat	circ.age.rodents	0.0501	-0.0127	0.1128	0.1191
rat	circ.age.eutherian	0.0799	0.0223	0.1375	0.0068
rat	circ.age.therian	0.0929	-0.0156	0.2014	0.0945
rhesus	genomic.length	-0.0389	-0.0531	-0.0248	0.0
rhesus	circ.age.primates	0.0376	0.0069	0.0684	0.0166
rhesus	circ.age.eutherian	0.1464	0.0927	0.2	0.0
rhesus	circ.age.therian	0.033	-0.0514	0.1174	0.4439
rhesus	dG.500	-0.0144	-0.0286	-0.0003	0.0459
human	genomic.length	-0.0332	-0.0453	-0.0212	0.0
human	circ.age.primates	0.0546	0.0265	0.0828	0.0001
human	circ.age.eutherian	0.0812	0.0334	0.129	0.0009
human	circ.age.therian	0.0727	-0.0023	0.1478	0.0579



Supplementary Figure 5: TE environment of dominant circRNAs | A: Opossum. The frequency of different TEs in flanking introns of dominant and randomly sampled circRNAs from the same hotspot was assessed. Sampling was repeated 1000 times for each dominant circRNA. The mean species overlap of the sampled circRNA was used to calculate the enrichment. Plotted is the log2-enrichment for different repeat families. An increase in color intensity reflects the mean number of TEs detected in the flanking introns of dominant circRNAs. Significance was estimated with a one-tailed and paired Fisher's t test. **B:** Rhesus. *Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.5.*

7 Curriculum Vitae and List of Publications

Franziska Gruhl

Home
Avenue des Figuiers 20, c/o Noth
1007 Lausanne
Email: franziska.gruhl@unil.ch

University
Center for Integrative Genomics
University of Lausanne
Quartier Sorge, 1015 Lausanne

Education

2013 - 2017 **Doctoral researcher, University of Lausanne, Switzerland**
Supervision by Ioannis Xenarios, Co-supervision Henrik Kaessmann
Expected thesis end in July 2017.

Thesis title: Recently active transposable elements provide insights into the evolution of mammalian circular RNAs

2010 - 2013 **MSc in Developmental Biology, University of Heidelberg, Germany**
Supervision by Joachim Wittbrodt, Centre for Organismal Studies (COS)

Thesis title: Morphological and molecular characterization of the Medaka (Oryzias latipes) digestive tract

2007 - 2010 **BSc in Biology, University of Heidelberg, Germany**
Supervision by Laurence Ettwiller, Institute of Zoology

Thesis title: Identification of co-regulating enhancer based on motif composition

Research Experience

Mar - Sep 2012 **Internship at University of Queensland Diamantina Institute, Australia**
Marcel Dinger, Role of lncRNAs in gene regulation

Jun- Dec 2012 **Research Associate at EMBL Heidelberg, Germany**
Detlev Arendt, Evolution of the bilaterian nervous system

Summer 2010 **Internship at Australian National University Canberra, Australia**
Iain Searle, Epigenetic regulation of root development in A. thaliana

2009 - 2011 **Research Associate at COS Heidelberg, Germany**
Laurence Ettwiller, Cis-regulatory elements in development of the vertebrate nervous system

Research Skills

Computational skills	Experienced in programming languages ruby, R, bash shell scripting; good knowledge of analysis tools for transcriptome analyses (e.g. cufflinks, cuffnorm, cuffquant, htseq, dexseq, bioconductor)
Statistical skills	good knowledge of statistical methods for biological data mining and transcriptome analysis
Molecular Biology	Molecular cloning and PCR techniques; immunohistochemical and histological staining techniques; cell proliferation analysis; in-vitro cell culture (mouse ES cells, conventional cell lines); lentivirus production; FACS analysis (plant cells); light and confocal microscopy; RNA-seq library preparation

Teaching and seminar organization

2013 - 2017	Organization of various seminars including internal and external speaker, organization of several sessions for the annual institutional retreat (e.g. Impact Factors and Publishing, Science Communication)
2013 - 2015	Tutor for R and statistics
2013 - 2014	Tutor for Molecular Evolution practical

Conferences

2015	Poster at BC2, Basel (Switzerland)
2015	Poster at SMBE, Vienna (Austria)
2012	Oral presentation at RNAi Australia (Australia)

Fellowships

2013 - 2017	PhD Fellowship by the Swiss Institute of Bioinformatics

Additional roles

2015 - 2017	PhD representative at the Institute

Publications

Accepted

* Authors contributed equally to the work

- Gloss B, Signal B, Cheetham S, **Gruhl F**, ... and Dinger ME
  "Dissecting developmental dynamics with high resolution temporal transcriptomics"
  Scientific Reports, 7(1), 6731. http://doi.org/10.1038/s41598-017-06110-5

- Aghaallaei N*, **Gruhl F***, Schaefer CQ, ... and Wittbrodt J, 2016
  "Identification, visualization and clonal analysis of intestinal stem cells in fish"
  Development, 143(19), 3470-3480. http://doi.org/10.1242/dev.134098

- Bell CC, Amaral PP, ..., **Gruhl F**, ... and Perkins AC, 2016
  "The Evx1/Evx1as gene locus regulates anterior-posterior patterning during gastrulation"
  Scientific Reports, 6, 26657. http://doi.org/10.1038/srep26657

- SIB Swiss Institute of Bioinformatics Members, 2016
  "The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases"
  Nucleic Acids Research, 44(D1), D27-37. http://doi.org/10.1093/nar/gkv1310

- Cheetham SW*, **Gruhl F***, Mattick JS and Dinger ME, 2013
  "Long noncoding RNAs and the genetics of cancer"
  British Journal of Cancer, 108(12), 2419-2425. http://doi.org/10.1038/bjc.2013.233

- ENCODE Project Consortium, 2012
  "An integrated encyclopedia of DNA elements in the human genome"
  Nature, 489(7414), 57-74. http://doi.org/10.1038/nature11247

- Mongin E, Auer TO, Bourrat F, **Gruhl F**, ... and Ettwiller L, 2011
  "Combining computational prediction of cis-regulatory elements with a new enhancer assay to efficiently label neuronal structures in the medaka fish"
  PloS One, 6(5), e19747. http://doi.org/10.1371/journal.pone.0019747

In preparation

- Mangul S, Yang HT, Strauli N, **Gruhl F**, ... and Zaitlen N
  "Dumpster diving in RNA-sequencing to find the source of every last read"
  Article is submitted to bioRxiv: `http://biorxiv.org/content/early/2016/05/13/053041`

Acknowledgements II

A PhD is never a straight path, but instead full of crossings and alternative roads - maybe a bit like a big tree that you have to climb. In the beginning, there are only a few branches and you kind of know which one to take. But then you have more and more branches and each one could lead you to the top. Sometimes you fall, sometimes you need to climb back, sometimes you need the help of others to figure out where to go. I would like to thank all the people who have been part in this journey of climbing my PhD tree.

My colleagues and institutes

The old lab in Lausanne - Thank you for having created this nice and relaxed working atmosphere. I've seen each of you leaving and every time it got a bit more grey making me realize how special it is to work in a lab with people like you.

The new lab in Heidelberg - Thank you for always having welcomed me when I came to Heidelberg, especially to Tania, Margarida and Evgeny who always tried to feed me with the newest information on what is going on and for always having an open ear and trying to make the time in Heidelberg as much fun as possible. All three of you have also discussed my project with me and it was great to have your input.

Vital-IT - Thank you for adopting me, the nice (and calm) atmosphere in the Mezzanine, all the interest in my project during the Vital-IT meetings as well as making sure that all the computational infrastructure works so well here!

The SIB - I remember the first SIB days I joined in 2013, seeing all PIs dancing the Gangdom style on the big stage or wherever else they found a spot and me thinking "I wanna be part of this sect" - well thanks for letting me be a part of it. It's been a pleasure to take all this excellent courses organized by the Trainee network, having met so many people from all the SIB groups, seeing how people work together and are trying to be part of something bigger! Thank you also to Vanessa, Véronique and Joceline for supporting me when I was having all these issues with the accident insurance. I don't know how I would have managed it without you.

The CIG - Thank you for providing this relaxed working atmosphere. It's been a pleasure to work at the institute and I'm thankful to all the people (the 3rd floor, Alex, Kostas, Julie, Tom, Olga, Simone, Patrick, ...), who have supported me especially in the last year while being here without my old lab.

My family and friends

Meine Familie - Danke an alle, an Mutti, Papa, Tommi, Jule und meine Oma und mein Opa für die ganze Unterstützung und Fürsorge in den letzten Jahren und eigentlich schon immer. Es ist gut zu wissen, dass man eine Familie hat, egal wo, zu der man immer kann, die Welt vergessen und einfach nur die Gedanken abschalten. Nicht jeder kann sich glücklich schätzen so starke Familienbände zu haben.

Iris - Simply thank you for being you! We have shared so many moments as friends and colleagues, be it while travelling together, while working late hours, simply while having a cup of coffee and exchanging the newest stories about each other's life. Not to think about each moment you were playing my second Mum, because I can be stubborn and naive like a little child.

Marieke - Well, my dear Dutch girl, thank you for all the moments we have shared together. Hiking, cooking at your place, making stupid jokes, thinking about the CIG, having drinks in the city, for all the laughter and thoughts we have exchanged. You have grown one of my closest friends here, and I'm grateful that we have met during the Christmas Party preparations for the 3rd floor - otherwise, who knows what would have happened with us?!

Kostya - So, originally there were some words in Russian supposed to be here, but texmaker crashes every time, and it seems to risky to me to continue trying just one day before handing in. Tak! I know we don't say thank you to each other, but sometimes it's good to do so. I'm happy to have you in my life, all your crazy ideas and deals (although I'm still not convinced we should try all of them). I'm thankful for every moment you make me laugh or I have to smile, because I hear you "stomping" up the stairs. Xpj-xrj!

Parkour Lausanne - Alors les traceurs et traceuses, merci pour être simplement magnifique! J'avais beaucoup de plaisir tracer avec vous et faire connaissance avec Lausanne dans une façon complémente différente. Merci à tout le monde de m'avoir montré le monde de Parkour. Je suis fière d'avoir vu comment vous vous êtes développés dans les dernières années. Tous les différents projets vous avez lancé comme la promotion du Parkour féminin, les rencontres avec des autres traceurs, les cours pour les gamins qu'on donne ensemble avec Nathy et Gaïth, et tous les prochains projets vous déjà imaginerez. Tellement, c'est bien de savoir qu'il y a des associations comme vous !

Christophe - Christophe, I think you have taught me a lot about myself and the important things in life! I always enjoy they time when we meet in Heidelberg or when you visit me in Lausanne and I'm grateful to have you as a friend. Merci beaucoup!

Mihai - Guten Morgen! Do you know that we have our 10 years anniversary this year? Yes, that much time has passed since we met in the Student's Elite Paradise in Heidelberg. We have shared a lot of experiences together, starting from the Pool Parties in winter, containern, different sports, geo cashing... You have been always there when I needed a friend, or simply again a couch in Heidelberg. Danke Mihai!

Marzia - Thank you Marzia for being such a good friend! Luckily, we both knew Marieke so that we could ran into each other at her house-warming party.

Margarida - Thank you for always being there and listening when I needed some advice in life and for always looking from the positive site on problems and making them see less severe. I truly admire this site of you!

Michael - Thank you for all the afternoons and working-home-from-the-city-evenings, all the conversations we had, that you trusted me with taking care of your snakes and spiders and that you always say your opinion and thoughts directly.

Hugo - Merci pour tout le temps qu'on a passé - soit s'entraîner ensemble au break dance, soit simplement discuter la vie. J'ai vraiment trouvé un bon ami avec toi.

Ma coloc - Alex et Nico, multilingual, always a bit chaotic - it's been a pleasure to share a flat with you. Merci and Danke!