Circular RNA repertoires are associated with evolutionarily young transposable elements

Franziska Gruhl[1,2‡], **Peggy Janich**[1,3§], **Henrik Kaessmann**[4*], **David Gatfield**[1*]

[1]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; [2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3]Krebsliga Schweiz, CH-3001 Bern, Switzerland; [4]Center for Molecular Biology of Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, Heidelberg, Germany

***For correspondence:** h.kaessmann@zmbh.uni-heidelberg.de (HK); david.gatfield@unil.ch (DG)

Present address: [†]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland; [‡]Krebsliga Schweiz, Bern, Switzerland

Abstract Circular RNAs (circRNAs) are found across eukaryotes and can function in post-transcriptional gene regulation. Their biogenesis through a circle-forming backsplicing reaction is facilitated by reverse-complementary repetitive sequences promoting pre-mRNA folding. Orthologous genes from which circRNAs arise, overall contain more strongly conserved splice sites and exons than other genes, yet it remains unclear to what extent this conservation reflects purifying selection acting on the circRNAs themselves. Our analyses of circRNA repertoires across five species representing three mammalian lineages (marsupials, eutherians: rodents, primates) reveal that surprisingly few circRNAs arise from orthologous exonic loci across different species. Even the circRNAs from the orthologous loci are associated with young, recently active and species-specific transposable elements, rather than with common, ancient transposon integration events. These observations suggest that many circRNAs emerged convergently during evolution – as a byproduct of splicing in orthologs prone to transposable element insertion. Overall, our findings argue against widespread functional circRNA conservation.

Introduction

First described more than forty years ago, circular RNAs (circRNAs) were originally perceived as a curiosity of gene expression, but they have gained significant prominence over the last 5-10 years (reviewed in ***Kristensen et al.*** (***2019***); ***Patop et al.*** (***2019***)). Large-scale sequencing efforts have facilitated the identification of thousands of individual circRNAs with specific expression patterns and, in some cases, specific functions (***Conn et al., 2015***; ***Du et al., 2016***; ***Hansen et al., 2013***; ***Piwecka et al., 2017***). CircRNA biogenesis occurs through so-called "backsplicing" events, in which an exon's 3' splice site is ligated onto an upstream 5' splice site of an exon on the same RNA molecule (rather than downstream, as in conventional splicing). Backsplicing occurs co-transcriptionally and is guided by the canonical splicing machinery (***Guo et al., 2014***; ***Ashwal-Fluss et al., 2014***; ***Starke et al., 2015***). It can be facilitated by complementary, repetitive sequences in the flanking introns (***Dubin et al., 1995***; ***Jeck et al., 2013***; ***Ashwal-Fluss et al., 2014***; ***Zhang et al., 2014***; ***Liang and Wilusz, 2014***; ***Ivanov et al., 2015***). Through intramolecular base-pairing and folding, the resulting hairpin-like structures can augment backsplicing over the competing, regular forward-splicing reaction. In most cases, backsplicing seems to be rather inefficient, given that circRNA expression levels are low in most tissues. For example, it has been estimated that about 60% of circRNAs exhibit expression levels of less than 1 FPKM (fragments per kilobase per million reads mapped) - a commonly applied cut-off below which genes are usually considered to not be robustly expressed (***Guo et al., 2014***). Due to their circular structure, circRNAs are protected from the activity of cellular exonucleases, which is thought to favour their accumulation to detectable steady-state levels and, together with the cell's proliferation history, presumably contributes to their complex spatiotemporal expression patterns (***Alhasan et al., 2015***; ***Memczak et al., 2013***; ***Bachmayr-Heyda et al., 2015***). Overall higher circRNA abundances have been reported for neural tissues (***Westholm et al., 2014***; ***Gruner et al., 2016***; ***Rybak-Wolf et al., 2015***) and during ageing (***Gruner et al., 2016***; ***Xu et al., 2018***; ***Cortés-López et al., 2018***).

CircRNAs are found in all eukaryotes (protists, fungi, plants, animals) (***Wang et al., 2014***). Moreover, it has been reported that circRNAs are frequently generated from orthologous genomic regions across species such as mouse, pig and human (***Rybak-Wolf et al., 2015***; ***Venøet al., 2015***), and that their splice sites have elevated conservation scores (***You et al., 2015***). In these studies, circRNA coordinates were transferred between species to identify "conserved" circRNAs. However, the analyses did not distinguish between potential selective constraints actually acting on the circRNAs themselves, from those preserving canonical splicing features of genes in which they are formed (so-called "parental genes"). A further obstacle to a thorough evolutionary understanding lies in the observation that while long introns containing reverse complementary repeats seem to be a conserved feature of circRNA parental genes, the reverse complementary repeat sequences as such undergo rapid evolutionary changes (***Zhang et al., 2014***; ***Rybak-Wolf et al., 2015***). Finally, concrete examples for experimentally validated, functionally conserved circRNAs are still scarce. At least in part, the reason may lie in the difficulty to specifically target circular vs. linear transcript isoforms in loss-of-function experiments; only recently, novel dedicated tools for such experiments have been developed (***Li et al., 2020***). At the moment, however, the prevalence of conserved and hence likely functional circRNAs remains overall unclear.

Here, we set out to investigate the origins and evolution of circRNAs as well as potentially associated selective pressures. To this end, we generated a comprehensive set of circRNA-enriched RNA sequencing (RNA-seq) data from five mammalian species and three organs. Our analyses unveil that circRNAs are typically generated from a distinct class of genes that share characteristic structural and sequence features. Notably, we discovered that circRNAs are flanked by species-specific and recently active transposable elements (TEs). Our findings support a model according to which the integration of TEs is preferred in introns of genes with similar genomic properties, thus facilitating circRNA formation as a byproduct of splicing around the same exons of orthologous genes across different species. Together, our work suggests that most circRNAs - even when occurring in orthologs of multiple species and comprising the same exons - do nevertheless not trace back to common ancestral circRNAs but emerged convergently during evolution, facilitated by independent TE insertion events.

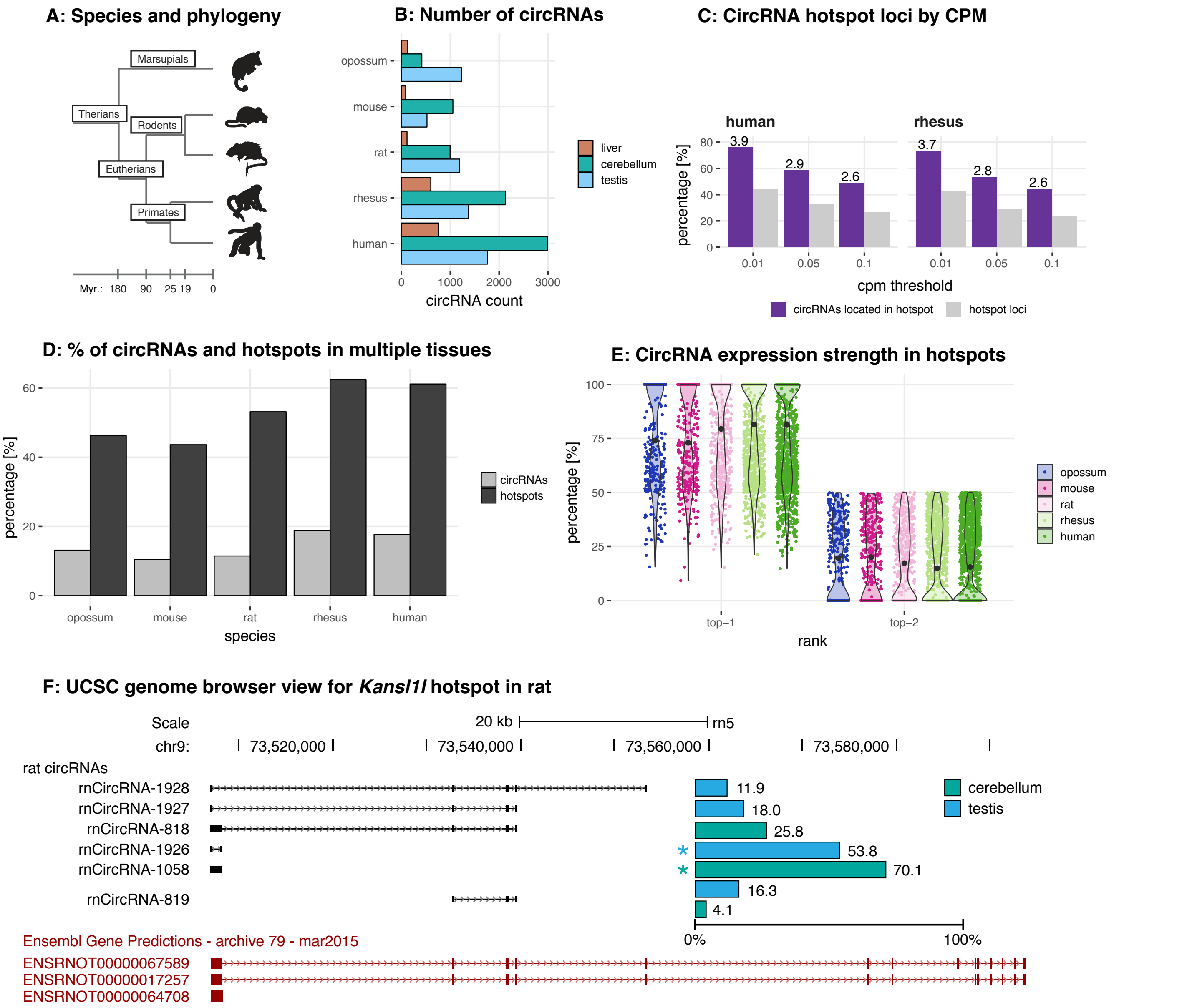
Results

A comprehensive circRNA dataset across five mammalian species

To explore the origins and evolution of circRNAs, we generated paired-end RNA-seq data for three organs (liver, cerebellum, testis) in five species (grey short-tailed opossum, mouse, rat, rhesus macaque, human) representing three mammalian lineages (marsupials; eutherians: rodents, primates) with different divergence times (**Figure 1A**, **Figure 1-Figure supplement 1A**, **Supplementary Table 1**). To enrich for circRNAs, samples were treated with exoribonuclease (RNase R) prior to library preparation and sequencing. Using a custom pipeline, we subsequently identified circRNAs from backsplice junction (BSJ) reads, estimated circRNA steady-state abundances, and reconstructed their isoforms (**Supplementary Table 2**, **Figure 1-Figure supplement 1B**, **Figure 1-Figure supplement 2**). In total, we identified 1,535 circRNAs in opossum, 1,484 in mouse, 2,038 in rat, 3,300 in rhesus macaque, and 4,491 circRNAs in human, with overall higher numbers in cerebellum, followed by testis and liver (**Figure 1B**, **Supplementary Table 3**). Detected circRNAs were generally small in size, overlapped with protein-coding exons, showed considerable tissue-specificity, and were flanked by large introns (**Figure 1-Figure supplement 3**).

The identification of circRNA heterogeneity and hotspot frequency is determined by sequencing depth and detection thresholds

A sizeable number of genes give rise to multiple, distinct circRNAs (***Venøet al., 2015***). Such “circRNA hotspots” are of particular interest as they may be enriched for genomic features that drive circRNA biogenesis. A previous hotspot definition applied a cutoff of at least 10 structurally different, yet overlapping circRNAs produced from a genomic locus (***Venøet al., 2015***). However, given that reaching a threshold of 10, or any other threshold, of detectable circRNA species for a given locus likely strongly depends on the sequencing depth and the applied CPM (counts per million) threshold, we compared circRNA hotspots identified at different CPM thresholds (0.1, 0.05 and 0.01 CPM). Moreover, to globally capture circRNA hotspot complexity, we considered genomic loci already as hotspots if they produced as a minimum two different, overlapping circRNAs at a given threshold. As expected, the number of hotspots - and the number of circRNAs these hotspots give rise to - strongly depend on the chosen CPM threshold (**Figure 1C** for human and rhesus macaque data; **Figure 1-Figure supplement 4** for other species). Thus, at 0.1 CPM only 16-27% of all detected circRNA loci are classified as hotspots. Decreasing the stringency to 0.01 CPM increases the proportion of hotspot loci to 32-45%. At the same time, the fraction of all circRNAs that originated from hotspots increased from 34-49% (0.1 CPM) to 59-76% (0.01 CPM), and the number of circRNAs per hotspot increased from 2 to 6. Together, these observations suggest that at lower CPM thresholds, it is in particular the number of circRNAs per locus that increases, whereas the effect on the number of detectable independent circRNA loci is smaller. Furthermore, we observed that in many cases the same hotspots produced circRNAs across multiple organs (**Figure 1D**), and that



there is usually one predominantly expressed circRNA per organ (**Figure 1E**). The *Kansl1l* hotspot locus is a representative example: it is a hotspot in rat, where it produces 6 different circRNAs (details in **Figure 1F**); it is also a hotspot in all other species (producing 8, 5, 7, and 6 different circRNAs in opossum, mouse, rhesus macaque and human, respectively; data not shown).

The substantial increase in circRNA heterogeneity with decreasing CPM, as well as the overall low expression levels of many circRNAs, raised the question to what extent the majority of detected circRNAs in this and other studies reflect a form of gene expression noise rather than functional transcriptome diversity.

Figure 1. A: Phylogenetic tree of species analysed in this study. CircRNAs were identified and analysed in five mammalian species (opossum, mouse, rat, rhesus macaque, human) and three organs (liver, cerebellum, testis). B: Total number of detected circRNAs across species and tissues. The total number of circRNAs for each species in liver (brown), cerebellum (green) and testis (blue). C: CircRNA hotspot loci by CPM (human and rhesus macaque). The graph shows, in grey, the proportion (%) of circRNA loci that qualify as hotspots and, in purple, the proportion (%) of circRNAs that originate from such hotspots, at three different CPM thresholds (0.01, 0.05, 0.1). The average number of circRNAs per hotspot is indicated above the purple bars. D. Number of circRNA hotspot loci found in multiple tissues. The graph shows the proportion (%) of circRNAs (light grey) and of hotspots (dark grey) that are present in at least two tissues. E. Contribution of top-1 and top-2 expressed circRNAs to overall circRNA expression from hotspots. The plot shows the contribution (%) that the two most highly expressed circRNAs (indicated as top-1 and top-2) make to the total circRNA expression from a given hotspot. For each plot, the median is indicated with a grey point. F. Example of the *Kansl1l* hotspot in rat. The proportion (%) for each detected circRNA within the hotspot and tissue (cerebellum = green, testis = blue) are shown. The strongest circRNA is indicated by an asterisk. rnCircRNA-819 is expressed in testis and cerebellum.

Figure 1–Figure supplement 1. Overview of the dataset and the reconstruction pipeline.

Figure 1–Figure supplement 2. Mapping summary of RNA-seq reads.

Figure 1–Figure supplement 3. General circRNA properties.

Figure 1–Figure supplement 4. CircRNA hotspot loci by CPM (opossum, mouse, rat).

CircRNAs formed in orthologous loci across species preferentially comprise constitutive exons

We therefore sought to assess the selective preservation – and hence potential functionality – of circRNAs. For each gene, we first collapsed circRNA coordinates to identify the maximal genomic locus from which circRNAs can be produced (**Figure 2A**). In total, we annotated 5,428 circRNA loci across all species (**Figure 2A**). The majority of loci are species-specific (4,103 loci; corresponding to 75.6% of all annotated loci), whereas there are only comparatively few instances where circRNAs arise from orthologous loci in the different species (i.e., from loci that share orthologous exons in corresponding 1:1 orthologous genes; **Figure 2A**). For example, only 260 orthologous loci (4.8% of all loci) give rise to circRNAs in all five species (**Figure 2A**). A considerable proportion of these shared loci also correspond to circRNA hotspots (opossum: 30.0%, mouse: 25%, rat: 32.3%, rhesus macaque: 44.6%, human: 60.4%). Thus, despite applying circRNA enrichment strategies in library preparation and lenient thresholds for computational detection, the number of potentially conserved orthologous circRNAs is surprisingly low.

PhastCons conservation scores are based on multiple alignments and known phylogenies, describing the conservation levels at single-nucleotide resolution (***Siepel et al., 2005***). To assess whether circRNA exons differed from non-circRNA exons in their conservation levels, we calculated phastCons scores for different exon types (circRNA exons, non-circRNA exons and UTR-exons). CircRNA exons showed higher phastCons scores in comparison to exons from the same genes that were not spliced into circRNAs (**Figure 2B**), which would be the expected outcome if purifying selection acted on functionally conserved circRNAs. However, other mechanisms may be relevant as well; constitutive exons, for example, generally exhibit higher conservation scores than alternative exons (***Modrek and Lee, 2003***; ***Ermakova et al., 2006***). We thus analysed exon features in more detail. First, the comparison of phastCons scores between exons of non-parental genes, parental genes and circRNAs revealed that parental genes were *per se* highly conserved (**Figure 2B**): 85-95% of the observed median differences between circRNA exons and non-parental genes could be explained by the parental gene itself. Next, we compared the usage of parental gene exons across organs (**Figure 2C**). We observed that circRNA exons are more frequently used in isoforms expressed in multiple organs than non-circRNA parental gene exons. Finally, we analysed the sequence composition at the splice sites, which revealed that GC amplitudes (i.e., the differences in GC content at the exon-intron boundary) are significantly higher for circRNA-internal exons than for parental gene exons that were located outside of circRNAs (**Figure 2D**).

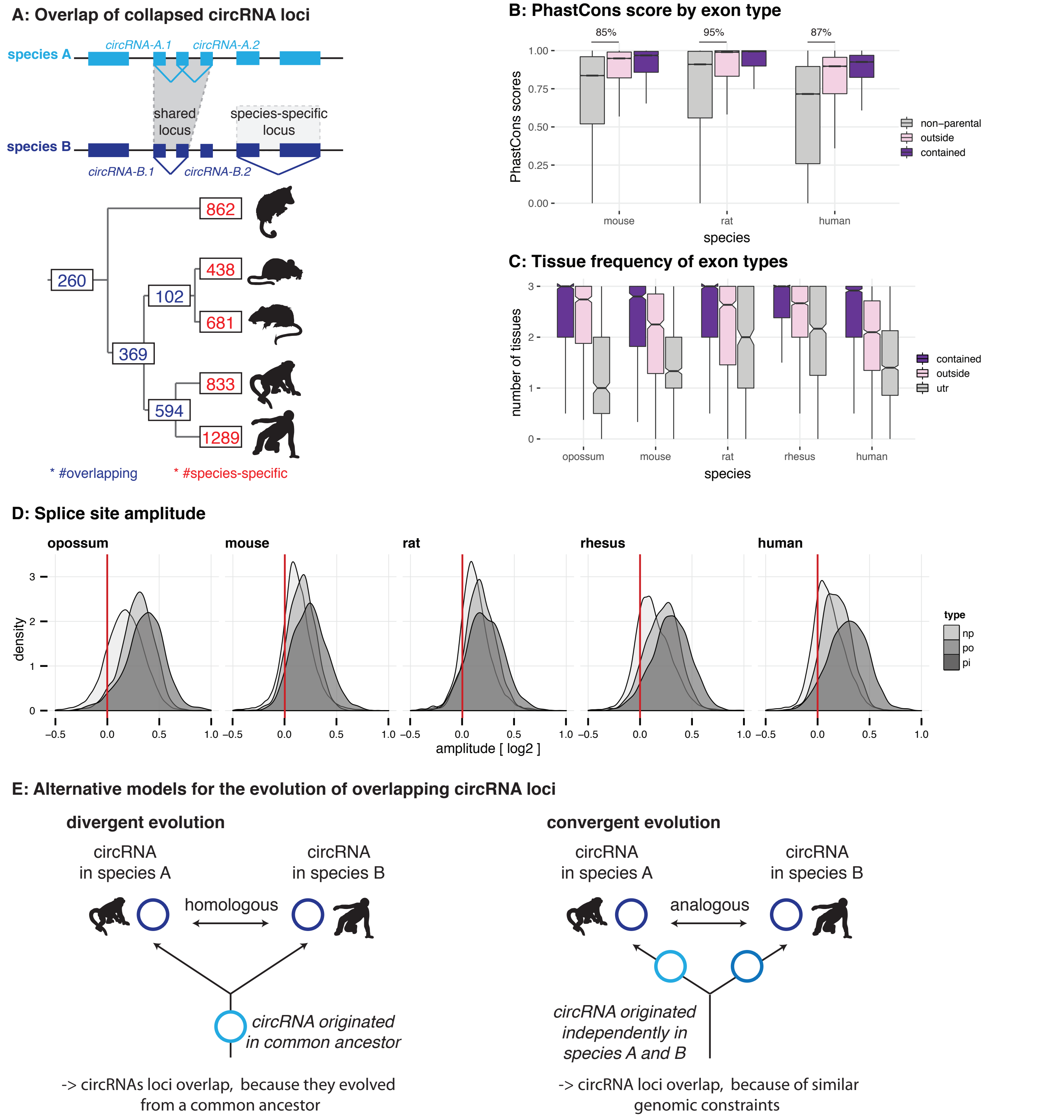
Collectively, these observations (i.e., increased phastCons scores, expression in multiple tissues, increased GC amplitudes) raise the question of whether the above “circRNA-specific” exon properties (**Figure 2B-D**) primarily reflect an enrichment for constitutive exons. Under this scenario, the supposed high conservation of circRNAs may not be directly associated with the circRNAs themselves, but with constitutive exons that the circRNAs contain. Together with the small proportion of circRNAs with shared (orthologous) locations across species (see above), this raises the possibility that circRNAs are overall not highly conserved and that many circRNAs “shared” across species (i.e., those arising from orthologous exonic loci) are actually not homologous. That is, rather than reflecting (divergent) evolution from common ancestral circRNAs (**Figure 2E, left panel**), circRNAs may frequently have emerged independently (convergently) during evolution in the lineages leading to the different species, thus potentially often representing “analogous” transcriptional traits (**Figure 2E, right panel**).

CircRNA parental genes are characterised by low GC content and high sequence repetitiveness

To explore whether convergent evolution played a role in the origination of circRNAs, we set out to identify possible structural and/or functional constraints that may establish a specific genomic environment (a “parental gene niche”) potentially favouring analogous circRNA production. To this end, we compared GC content and sequence repetitiveness of circRNA parental vs. non-parental genes.

GC content is an important genomic sequence characteristic associated with distinct patterns of gene structure, splicing and function (***Amit et al., 2012***). We realised that the increased GC amplitude at circRNA exon-intron boundaries that we noted above (**Figure 2D**), was mainly caused by a local decrease of intronic GC content rather than an increase in exonic GC content (**Supplementary Table 4**, **Figure 2-Figure supplement 2**). We hence hypothesised that GC content may be a means of discriminating parental from non-parental genes. We grouped genes into five categories from low (L) to high (H) GC content (isochores; L1 <37%, L2 37-42%, H1 42-47%, H2 47-52% and H3 >52% GC content) (**Figure 3A**). Coding genes in rhesus macaque and human are characterised by a bimodal GC content distribution (see peaks in L2 and H3 for non-parental genes). By contrast, the two rodents displayed a unimodal distribution (peak in H1), whereas opossum coding genes were generally GC-poor (in agreement with ***Galtier and Mouchiroud*** (***1998***); ***Mikkelsen et al.*** (***2007***)). Notably, circRNA parental genes showed a distinctly different distribution than non-parental genes and a consistent pattern across all five species, with the majority of genes (82-94% depending on species) distributing to the GC-low gene groups, L1 and L2 (**Figure 3A**).

We next analysed intron repetitiveness – a structural feature that has previously been asso-



ciated with circRNA biogenesis. We used megaBLAST to align all annotated coding genes with themselves to identify regions of complementarity in the sense and antisense orientations of the gene (reverse complement sequences, RVCs) (***Ivanov et al., 2015***). We then compared the level of self-complementarity between parental and non-parental genes within the same isochore (i.e., per gene group with the same GC content), given that self-complementarity generally shows neg-

Figure 2. Evolutionary properties of circRNAs. A: CircRNA loci overlap between species. Upper panel: Schematic representation of the orthology definition used in our study. CircRNAs were collapsed for each gene, and coordinates were lifted across species. Lower panel: Number of circRNA loci that are species-specific (red) or circRNAs that arise from orthologous exonic loci of 1:1 orthologous genes (i.e., circRNAs sharing 1:1 orthologous exons) across lineages (purple) are counted. We note that in the literature, other circRNA "orthology" definitions can be found, too. For example, assigning circRNA orthology simply based on parental gene orthology implies calling also those circRNAs "orthologous" that do not share any orthologous exons, which directly argues against the notion of circRNA homology; that is, a common evolutionary origin (see **Figure 2-Figure supplement 1**). Overall, the orthology considerations we applied largely follow the ideas sketched out in ***Patop et al.*** (***2019***). B: Distribution of phastCons scores for different exon types. PhastCons scores were calculated for each exon using the conservation files provided by ensembl. PhastCons scores for non-parental exons (grey), exons in parental genes, but outside of the circRNA (pink) and circRNA exons (purple) are plotted. The difference between circRNA exons and non-parental exons that can be explained by parental non-circRNA exons is indicated above the plot. C: Mean tissue frequency of different exon types in parental genes. The frequency of UTR exons (grey), non-UTR exons outside of the circRNA (pink) and circRNA exons (purple) that occur in one, two or three tissues was calculated for each parental gene. D: Distribution of splice site amplitudes for different exon types. Distribution of median splice site GC amplitude (log2-transformed) is plotted for different exon types (np = non-parental, po = parental, but outside of circRNA, pi = parental and inside circRNA). Red vertical bars indicate values at which exon and intron GC content would be equal E: Different evolutionary models explaining the origins of overlapping circRNA loci.

Figure 2–Figure supplement 1. CircRNA loci overlap between species.

Figure 2–Figure supplement 2. Amplitude correlations.

ative correlations with GC-content. This analysis revealed a stronger level of self-complementarity in sense and antisense for parental genes than for non-parental genes from the same isochore (**Figure 3B**).

CircRNA parental genes may also show an association with specific functional properties. Using data from three human cell studies (***Steinberg et al., 2015***; ***Pai et al., 2012***; ***Koren et al., 2012***), our analyses revealed that circRNA parental genes are biased towards early replicating genes, showed higher steady-state expression levels, and are characterised by increased haploinsufficiency scores (**Figure 3-Figure supplement 1**). Collectively, we conclude that circRNA parental genes exhibit not only distinct structural features (low GC content, high repetitiveness), but also specific functional properties associated with important roles in human cells.

Among the multiple predictors of circRNA parental genes, low GC content distinguishes circRNA hotspots

The aforementioned analyses established that circRNA parental genes possess distinct sequence, conservation and functional features. Using linear regression analyses, we next sought to determine which of these properties constitute the main predictors of parental genes. Our model used parental vs. non-parental gene as the response variable and several plausible explanatory variables (i.e., GC content, exon and transcript counts, genomic length, number of repeat fragments in sense/antisense, expression level, phastCons score, tissue specificity index). After training the model on a data subset (80%), circRNA parental gene predictions were carried out on the remainder of the dataset (20%) (see **Material and Methods** for more information). Notably, predictions occurred with high precision (accuracy 72-79%, sensitivity of 75%, specificity 71-79% across all species) and uncovered several significantly associated features (**Table 1**, **Supplementary Table 5**, **Figure 3-Figure supplement 2**). Consistently for all species, the main parental gene predictors are low GC content (log-odds ratio -1.84 to -0.72) and increased number of exons in the gene (log-odds

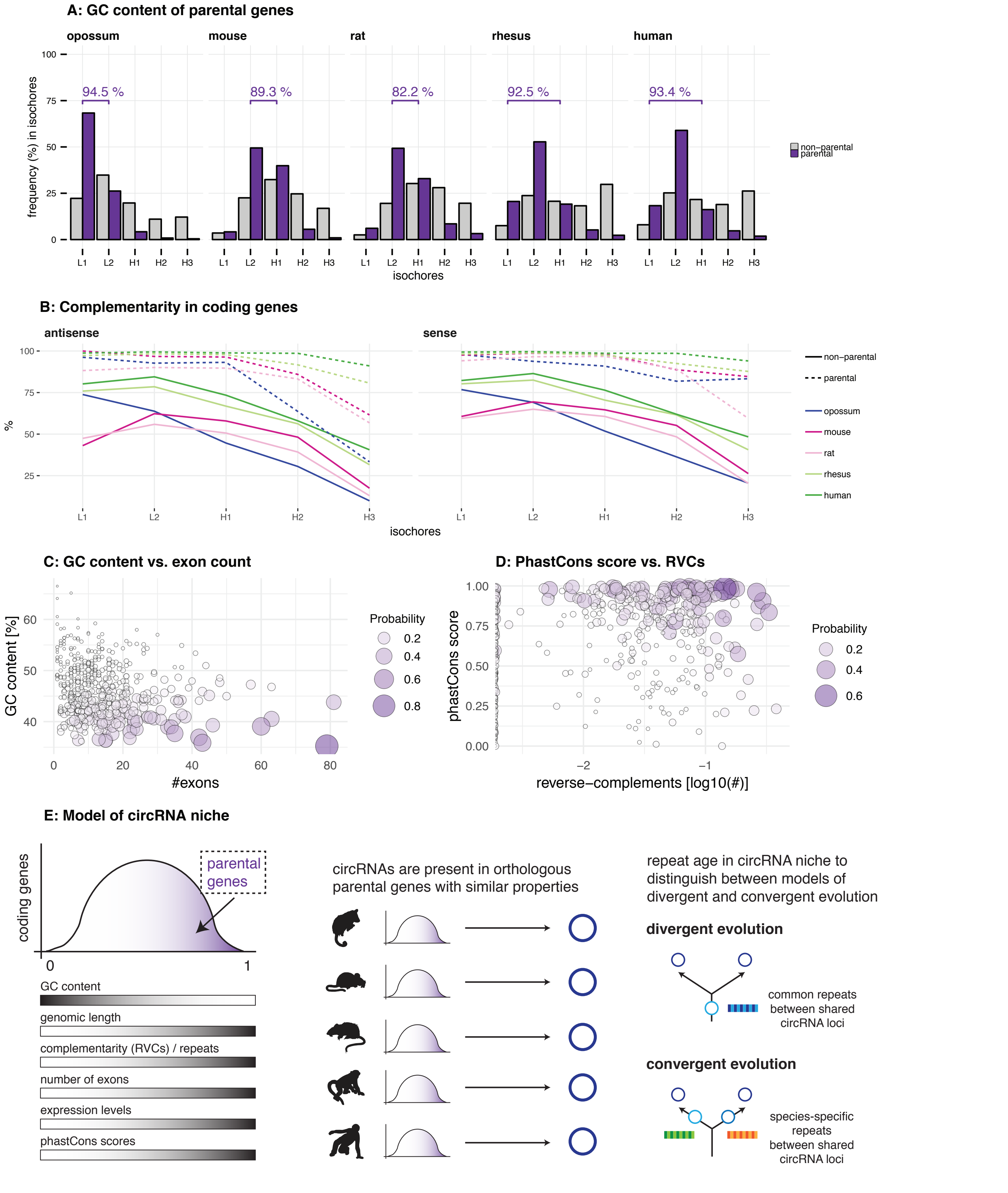

A: GC content of parental genes
opossum
mouse
rat
rhesus
human
frequency (%) in isochores
100
75
50
25
0
94.5 %
89.3 %
82.2 %
92.5 %
93.4 %
L1
L2
H1
H2
H3
isochores
non–parental
parental
B: Complementarity in coding genes
antisense
sense
%
100
75
50
25
L1
L2
H1
H2
H3
isochores
non-parental
parental
opossum
mouse
rat
rhesus
human
C: GC content vs. exon count
GC content [%]
60
50
40
0
20
40
60
80
#exons
Probability
0.2
0.4
0.6
0.8
D: PhastCons score vs. RVCs
phastCons score
1.00
0.75
0.50
0.25
0.00
−2
−1
reverse−complements [log10(#)]
Probability
0.2
0.4
0.6
E: Model of circRNA niche
coding genes
parental genes
0
1
GC content
genomic length
complementarity (RVCs) / repeats
number of exons
expression levels
phastCons scores
circRNAs are present in orthologous parental genes with similar properties
repeat age in circRNA niche to distinguish between models of divergent and convergent evolution
divergent evolution
common repeats between shared circRNA loci
convergent evolution
species-specific repeats between shared circRNA loci

Figure 3. A: GC content of parental genes. Coding genes were classified into L1-H3 based on their GC content, separately for non-parental (grey) and parental genes (purple). The percentage of parental genes in L1-L2 (opossum, mouse, rat) and L1-H1 (rhesus macaque, human) is indicated above the respective graphs. B: 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 against its isochore. C-D: Examples of parental gene predictors for linear regression models. A generalised linear model (GLM) was fitted to predict the probability of the murine coding gene to be parental, whereby x- and y-axis represent the strongest predictors. Colour and size of the discs correspond to the p-values obtained for 500 genes randomly chosen from all mouse coding genes used in the GLM. E. Model of circRNA niche.

Figure 3–Figure supplement 1. Replication time, gene expression steady-state levels and GHIS of human parental genes.

Figure 3–Figure supplement 2. Distribution of prediction values for non-parental and parental circRNA genes.

Figure 3–Figure supplement 3. Validation of parental gene GLM on Werfel *et al.* dataset.

ratio 0.30 to 0.45). Furthermore, increased genomic length (log-odds ratio 0.17 to 0.26) and an increased proportion of reverse-complementary areas (repeat fragments) within the gene (log-odds ratio 0.20 to 0.59), increased expression levels (log-odds ratio 0.25 to 0.38) and higher phastCons scores (log-odds ratio 0.45 to 0.58) are also positively associated with circRNA production (**Table 1**, **Figure 3C-D**, **Supplementary Table 5**). Notably, these circRNA parental gene predictors were not restricted to our datasets but could be deduced from independent circRNA datasets as well. Thus, the analysis of mouse and human heart tissue data (***Werfel et al., 2016***) revealed the same properties; that is, circRNA parental genes are characterised by low GC content, they were exon-rich, and they showed enrichment for repeats (**Figure 3-Figure supplement 3**). Moreover, our linear regression models performed with comparable accuracy (74%), sensitivity (75%) and specificity (74%) to predict parental genes in the independent human and mouse data. We therefore conclude that the identified properties likely represented generic characteristics of circRNA parental genes that are suitable to distinguish them from non-parental genes.

A substantial amount of circRNAs are formed from circRNA hotspots (**Figure 1C**). We therefore asked whether among the distinct genomic features that our regression analysis identified as characteristic of parental genes, some would be suitable to further distinguish hotspots. First, we assessed whether hotspots were more likely to be shared between species than parental genes producing only a single circRNA isoform. Notably, the applied regression model did not only detect a positive correlation between the probability of a parental gene to be a hotspot and having orthologous parental genes in multiple species, but log-odds ratios increased with the distance and number of species across which the hotspot was shared (e.g., mouse: 0.29 for shared within rodents, 0.67 for shared with eutherian species and 0.72 for shared within therian species; **Supplementary Table 6**). Finally, we interrogated whether a particular feature would be able to specify circRNA hotspots among parental genes. A single factor, low GC content, emerged as a consistent predictor for circRNA hotspots among all circRNA-generating loci (**Supplementary Table 7**). Not surprisingly, the predictive power was lower than that of the previous models discriminating parental vs. non-parental genes, which had identified low GC content as well. These findings imply that hotspots emerge across species in orthologous loci that offer similarly favourable conditions for circRNA formation, including low GC content. Of note, the increased number of circRNAs that become detectable when CPM thresholds are lowered (see above, **Figure 1C**), is also in agreement

Table 1. A generalised 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). The model was trained on 80% of the data (scaled values, cross-validation, 1000 repetitions). Only the best predictors were kept and then used to predict probabilities for the remaining 20% of data points (validation set, shown in table). Genomic length, number of exons and GC content are based on the respective ensembl annotations; number of repeats in antisense and sense orientation to the gene was estimated using the RepeatMasker annotation, phastCons scores taken from UCSC (not available for opossum and rhesus macaque) and expression levels and the tissue specificity index based on (***Brawand et al., 2011***). An overview of all log-odds ratios and p-values calculated in the validation set of each species is provided in the table, further details can be found in **Supplementary Table 5**. *Abbreviations: md = opossum, mm = mouse, rn = rat, rm = rhesus macaque, hs = human. Significance levels: '***' < 0.001, '**' < 0.01, '*' < 0.05, 'ns' >= 0.05.*

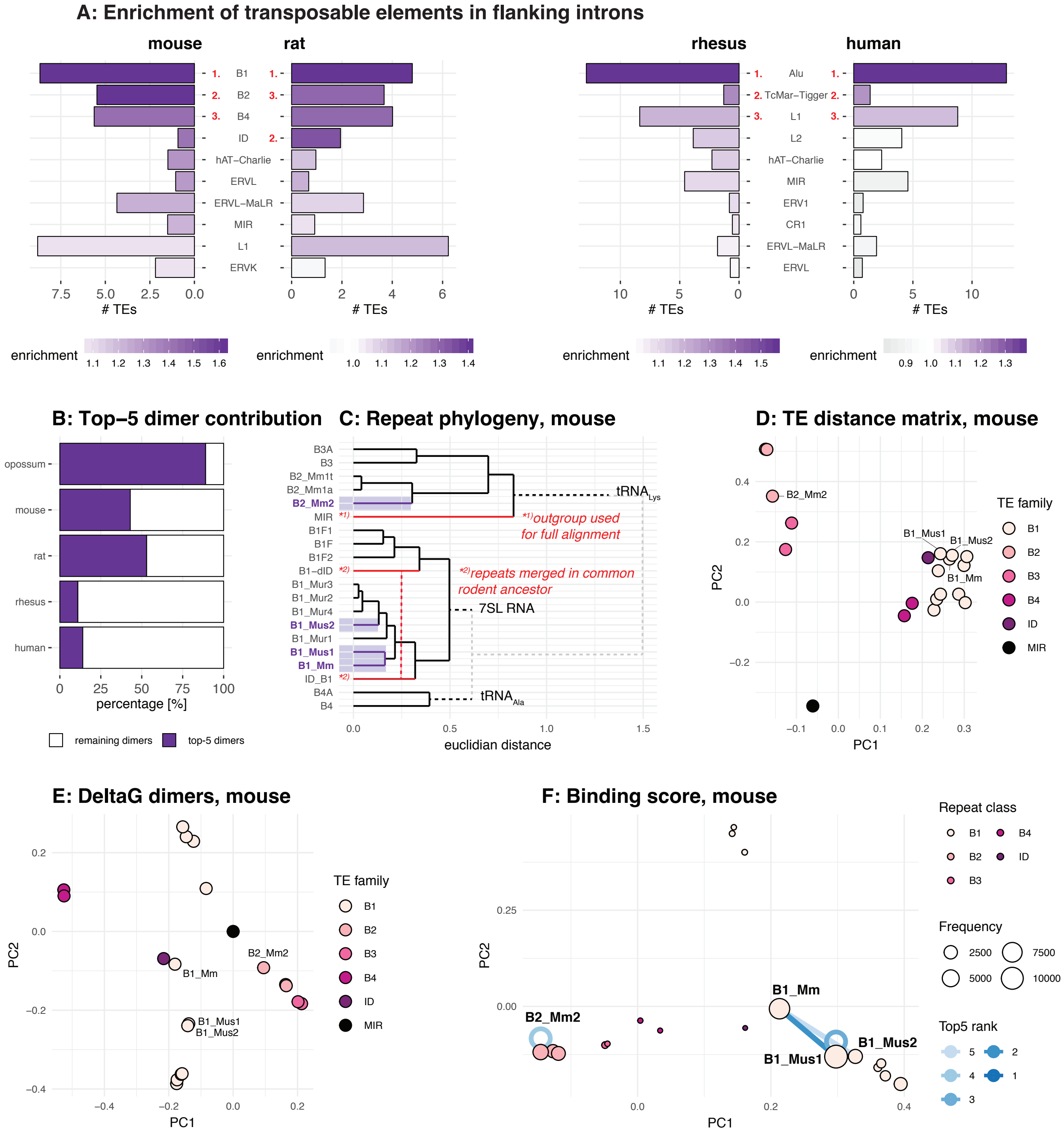
Predictor	Log-odds range (significance)	Species with significant predictor
Genomic gene length (bp)	rn: 0.26 (***) rm: 0.17 (***) hs: 0.26 (***) md, mm: ns	rn, rm, hs
Number of exons	md: 0.45 (***) mm: 0.38 (***) rn: 0.30 (***) rm: 0.42 (***) hs: 0.32 (***)	md, mm, rn, rm, hs
GC content	md: -1.84 (***) mm: -1.09 (***) rn: -0.72 (***) rm: -1.44 (***) hs: -1.42 (***)	md, mm, rn, rm, hs
Repeat fragments (antisense)	md: 0.28 (**) mm: 0.20 (**) rm: 0.59 (***) rn, hs: ns	md, mm, rm
Repeat fragments (sense)	hs: 0.58 (***) md, mm, rn, rm: ns	hs
PhastCons scores	mm: 0.58 (***) rn: 0.51 (***) hs: 0.45 (***)	mm, rn, hs
Mean expression levels	md: 0.34 (**) rm: 0.38 (***,) hs: 0.25 (**) mm, rn: ns	md, rm, hs
Tissue specificity index	md, mm, rn, rm, hs: ns	-

with the sporadic formation of different circRNAs whenever genomic circumstances allow for it.

Collectively, our analyses thus reveal that circRNA parental genes are characterised by a set of distinct features: low GC content, increased genomic length and number of exons, higher expression levels and increased phastCons scores (**Figure 3E**). These features were detected independently across species, suggesting the presence of a unique, syntenic genomic niche in which circRNAs can be produced ("circRNA niche"). While helpful to understand the genomic context of circRNA production, these findings do not yet allow distinguishing between the two alternative models of divergent and convergent circRNA evolution (**Figure 2E**). However, we reasoned that this aim would be in reach if we better understood the evolutionary trajectory and timeline that leads to the emergence of the circRNAs. Conceivably, the identified feature "complementarity and repetitiveness" of the circRNA niche might give access to this time component. Previous studies have associated repetitiveness with an over-representation of small TEs – such as primate Alu elements or the murine B1 elements – in circRNA-flanking introns; these TEs may facilitate circRNA formation by providing RVCs that are the basis for intramolecular base-pairing of nascent RNA molecules (***Ivanov et al., 2015***; ***Jeck et al., 2013***; ***Zhang et al., 2014***; ***Wilusz, 2015***; ***Liang and Wilusz, 2014***). Interestingly, while the biogenesis of human circRNAs has so far been mainly associated with the primate-specific group of Alu elements, a recent study has highlighted several circRNAs that rely on the presence of mammalian MIR elements (***Yoshimoto et al., 2020***). A better understanding of the evolutionary age of TEs in circRNA-flanking introns could thus provide important insights into the modes of circRNA emergence; that is, the presence of common (i.e., old) repeats would point towards divergent evolution of circRNAs from a common circRNA ancestor, whereas an over-representation of species-specific (i.e., recent) repeats would support the notion of convergent circRNA evolution (**Figure 3E**).

CircRNA flanking introns are enriched in species-specific TEs

To assess potential roles of TEs in circRNA evolution, we first investigated the properties and composition of the repeat landscape relevant for circRNA biogenesis - features that have remained poorly characterised so far - harnessing our cross-species dataset. As a first step, we generated for each species a background set of "control introns" from non-circRNA genes that were matched to the circRNA flanking introns in terms of length distribution and GC content. We then compared the abundance of different repeat families within the two intron groups. In all species, TEs belonging to the class of small, interspersed nuclear elements (SINEs) are enriched within the circRNA flanking introns as compared to the control introns. Remarkably, the resulting TE enrichment profiles were exquisitely lineage-specific, and even largely species-specific (**Figure 4A**). In mouse, for instance, the order of enrichment is from the B1 class of rodent-specific B elements (strongest enrichment and highest frequency of >7.5 TEs per flanking intron) to B2 and B4 SINEs. In rat, B1 (strong enrichment, yet less frequent than in mouse) is followed by ID (Identifier) elements, which are a family of small TEs characterised by a recent, strong amplification history in the rat lineage (***Kim et al., 1994***; ***Kim and Deininger, 1996***); B2 and B4 SINEs only followed in 3rd and 4th position. In rhesus macaque and human, Alu elements are the most frequent and strongly enriched TEs (around 14 TEs per intron), consistent with the known strong amplification history in the common primate ancestor (reviewed in ***Batzer and Deininger*** (***2002***)) (**Figure 4A**). The opossum genome is known for its high number of TEs, many of which may have undergone a very species-specific amplification pattern (***Mikkelsen et al., 2007***), which is reflected in the distinct opossum enrichment profile (**Figure 4-Figure supplement 1**).



As pointed out above, TEs are relevant for circRNA formation because they can provide the RVCs that are the basis for intramolecular base-pairing of nascent RNA molecules (***Ivanov et al., 2015***; ***Jeck et al., 2013***; ***Zhang et al., 2014***; ***Wilusz, 2015***; ***Liang and Wilusz, 2014***). Folding of the pre-mRNA

Figure 4. A: Enrichment of TEs in flanking introns for mouse, rat, rhesus macaque and human. The number of TEs was quantified in both intron groups (circRNA flanking introns and length- and GC-matched control introns). Enrichment of TEs is represented by colour from high (dark purple) to low (grey). The red numbers next to the TE name indicate the top-3 enriched TEs in each species. B: Top-5 dimer contribution. The proportion of top-5 dimers (purple) to the remaining dimers (white) in flanking introns is shown. C: Phylogeny of mouse TEs. Clustal-alignment based on consensus sequences of TEs. Most recent TEs are highlighted. D: PCA for distance matrix of mouse TE families. PCA is based on the clustal-alignment distance matrix for the reference sequences of all major SINE families in mouse with the MIR family used as an outgroup. TEs present in the top-5 dimers are labelled. E: PCA based on deltaG for mouse TE families. PCA is based on the minimal free energy (deltaG) for all major SINE families in mouse with the MIR family used as an outgroup. TEs present in the top-5 dimers are labelled. F: PCA for binding score of mouse dimers. PCA is based on a merged and normalised score, taking into account binding strength (=deltaG) and phylogenetic distance. Absolute frequency of TEs is visualised by circle size. TEs present in the five most frequent dimers (top-5) are highlighted by blue lines connecting the two TEs engaged in a dimer (most frequent dimer in dark blue = rank 1). If the dimer is composed of the same TE family members, the blue line loops back to the TE (= blue circle).

Figure 4–Figure supplement 1. Enrichment of transposable elements in flanking introns for opossum.

Figure 4–Figure supplement 2. PCA and phylogeny of opossum, rat, rhesus macaque and human repeat dimers.

into a hairpin secondary structure with a paired RNA stem (formed by the flanking introns via the dimerised RVCs) and an unpaired loop region (carrying the future circRNA) leads to a configuration that is favourable for circRNA formation because it brings backsplice donor and acceptor sites into close proximity. In order to serve as efficient RVCs via this mechanism, TEs will need to fulfil certain criteria, and the dimerisation potential will likely depend on TE identity, frequency, and position. Moreover, while two integration events involving the same TE (in reverse orientation) will lead to an extended RVC stretch, different transposons from the same TE family also still share varying degrees of sequence similarity that depend on their phylogenetic distance. The sequence differences that have evolved might compromise the base-pairing potential. To cover the dimerisation potential of the TE landscape in a comprehensive fashion, we deemed it vital to calculate the actual binding affinities between the dimerising sequences. As described below, we thus established a binding score that would account for this variety of factors influencing dimer formation and that would allow us to identify the TEs representing the most likely drivers of circRNA formation.

First, we noted that, similar to TEs overall (**Figure 4A**), RVCs were also enriched in SINE TEs (**Figure 4B**). Moreover, in some species, relatively few specific dimers represented the majority of all predicted dimers (i.e., top-5 dimers accounted for 89% of all dimers in flanking introns in opossum, 43% in mouse, 53% in rat, 11% in rhesus and 14% in human). We further realised that the phylogenetic distance between different TEs in a species was inadequate to categorise them with regard to their dimer potential; as shown for mouse (**Figure 4C-D**), phylogenetic age only separated large subgroups, but not TEs of the same family whose sequences have diverged by just a few nucleotides. By contrast, classification by binding affinities creates more precise, smaller subgroups that lack, however, the information on phylogenetic age (**Figure 4E**). Therefore, we devised a binding score that integrates both phylogeny (age) and binding affinity information (see **Material and Methods**). Principal component analysis (PCA) showed that it efficiently separated different TE families and individual family members, with PC1 and PC2 of the binding score explaining approximately 76% of observed variance (**Figure 4F**; **Figure 4-Figure supplement 2**). Moreover, this analysis suggests that the most frequently occurring dimers (top-5 dimers are depicted as blue connecting lines in **Figure 4F**) are formed by recently active TE family members. In mouse, an illustrative example are the dimers formed by the B1_Mm, B1_Mus1 and B1_Mus2 elements (**Figure 4F**), which are among the most recent (and still active) TEs in this species (**Figure 4C**). Across species, our analyses allowed for the same conclusions. For example, the dominant dimers in rat were precisely the recently amplified ID elements, and not the more abundant (yet older in their amplification history) B1 family of TEs (**Figure 4-Figure supplement 2B**) (***Kim et al., 1994***; ***Kim and Deininger, 1996***). In opossum, the most prominent dimers consisted of opossum-specific SINE1 elements, which are similar to the Alu elements in primates, but possess an independent origin (**Figure 4-Figure supplement 2A**) (***Gu et al., 2007***). Finally, dimer composition within the primate lineage was relatively similar, probably due to the high amplification rate of AluJ and AluS/Z elements in the common primate ancestor and relatively recent divergence time of macaque and human (**Figure 4-Figure supplement 2C-D**) (***Batzer and Deininger, 2002***).

In conclusion, the above analyses of RVCs revealed that dimer-forming sequences in circRNA flanking introns were most frequently composed of recent, and often currently still active, TEs. Therefore, the dimer repertoires were specific to the lineages (marsupials, rodents, primates) and/or even – as most clearly visible within the rodent lineage – species-specific.

Flanking introns of circRNA loci shared across species are enriched in evolutionarily young TEs

We next compared the dimer composition of the two groups of introns, namely those that flanked circRNA loci whose exonic locations are in common between species and those that flanked species-specific circRNA loci. For this analysis – aimed at finally resolving the extent to which circRNA loci shared across species evolved from a common ancestor or independently from each other – we took into account the degradation rate (milliDiv, see hereafter), frequency, enrichment and age of the dimers. Briefly, the RepeatMasker annotations (***Smit et al., 2013***) (http://repeatmasker.org; see **Material and Methods** for more details) provide a quantification of how many “base mismatches in parts per thousand” have occurred between each specific repeat copy in its genomic context and the repeat reference sequence. This deviation is expressed as the milliDiv value. Thus, a high milliDiv value implies that a repeat is strongly degraded, typically due to its age (the older the repeat, the more time its sequence has had to diverge). Low milliDiv values suggest that the repeat is younger (i.e., it had less time to accumulate mutations) or that purifying selection prevented the accumulation of mutations. Using this rationale, we explored degradation rates for the top-5 dimers extracted in each species from the ensemble of parental genes, and then compared the milliDiv values associated with orthologous genes giving rise to circRNAs in multiple species (shared parental genes) to those for species-specific parental genes. Notably, dimers detected in shared parental genes are generally less degraded than those in species-specific parental genes (**Figure 5A**). In rat, for example, median milliDiv values for the dimers involving young TE classes (ID_Rn1+ID_Rn1, ID_Rn1+ID_Rn2 and ID_Rn2+ID_Rn2) range from 21 to 42.5 for shared parental genes and 26 to 43.5 for species-specific parental genes, with the differences between shared and species-specific parental genes all being statistically significant (**Figure 5A, left panel**). By contrast, no significant milliDiv differences were found in the case of dimers involving older repeats (BC1_Rn+ID_Rn1 and BC1_Rn+ID_Rn2); thus, their degradation rates are comparable between shared and species-specific parental genes. The human data (**Figure 5A, right panel**) and that from opossum, mouse and macaque (**Figure 5-Figure supplement 1A**) revealed similar trends. For example, differences in degradation rates between shared and human-specific parental genes were observed for the dimers containing younger repeats such as AluSx1+AluY or AluSx+AluY (**Figure 5A, right panel**). In conclusion, these analyses reveal that flanking introns of circRNAs are enriched in TEs with rather species-specific integration and amplification rates, consistent with the idea of convergent circRNA evolution driven by independent TE insertion events in orthologous genomic loci.

Low degradation rates could indicate that specific dimers are particularly important for the production of functional circRNAs. For example, Alu elements, which the above dimer analyses identified as important in human and rhesus macaque, are common to the primate lineage, and it would be conceivable that the circRNA loci shared between both species emerged through TE integration in a common primate ancestor and were subsequently preserved by purifying selection. Alternatively, differences in degradation rates may simply reflect the evolutionary age of integration events. In that case, we would predict that even though the circRNA parental genes are shared between species, the enriched dimers would nevertheless stem from recent, independent integration events, rather than from ancestral, shared integration events. This scenario could occur if the circRNA-producing genes were to act as “transposon sinks” that are prone to insertions of active repeats due to specific features related to their sequence or structural architecture. To explore this idea, we examined in greater detail the dimers in shared and species-specific parental genes. As in our above analyses, we first created specific “dimer lists”, this time restricted to the two groups of parental genes (shared/species-specific circRNA loci); using the top-100 most and least enriched dimers, we compared the enrichment factors and mean age (categorised for simplicity into four groups: 1 = species-specific, 2 = lineage-specific, 3 = eutherian, 4 = therian). The analysis revealed that the most enriched and most frequent dimers are consistently formed by the youngest elements in both groups of genes, and that the frequency distribution of the top-100 dimers was significantly different between species (see **Figure 5B** for mouse and rhesus macaque; other species in **Figure 5-Figure supplement 1B**). In rhesus macaque, for example, the most frequent dimers included the Alu element AluYra, which is characteristic for this species and absent from the human lineage. A representative example for such a shared circRNA-generating locus with young, species-specific repeats is the *Akt3* locus (**Figure 5C**). Although *Akt3* circRNAs are shared between human (upper panel), mouse (middle panel) and opossum (lower panel), the dimer landscapes (top-5 dimers are highlighted in the figure) are entirely specifies-specific.

Taken together, we conclude that circRNAs are preferentially formed from loci that have acquired TEs in recent evolutionary history. Such recent transposition events involved TEs that have a higher degree of species-specificity than evolutionarily older TEs. Importantly, even in the case of genomic loci whose capacity to generate circRNAs was shared across species, the actual repeat landscapes revealed that they had acquired their TEs in evolutionarily recent times, as judged from repeat degradation rates and age. Overall, these findings support a model according to which circRNAs are analogous, rather than homologous features of loci that have increased propensity of attracting TEs, likely due to particular genomic features such as their GC content.

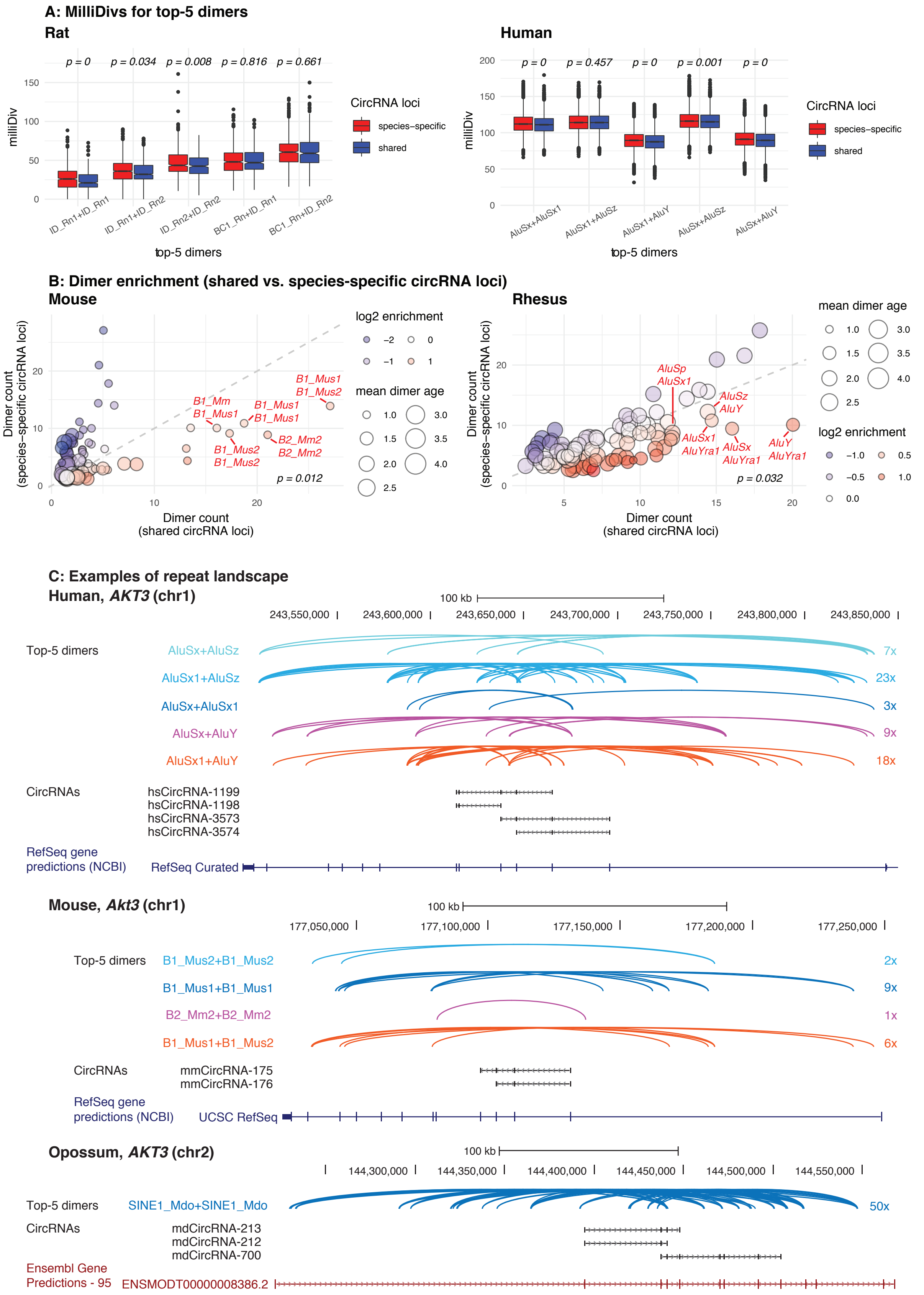

A: MilliDivs for top-5 dimers
Rat
p = 0
p = 0.034
p = 0.008
p = 0.816
p = 0.661
milliDiv
150
100
50
0
ID_Rn1+ID_Rn1
ID_Rn1+ID_Rn2
ID_Rn2+ID_Rn2
BC1_Rn+ID_Rn1
BC1_Rn+ID_Rn2
top-5 dimers
CircRNA loci
species–specific
shared
Human
p = 0
p = 0.457
p = 0
p = 0.001
p = 0
milliDiv
200
150
100
50
0
AluSx+AluSx1
AluSx1+AluSz
AluSx1+AluY
AluSx+AluSz
AluSx+AluY
top-5 dimers
CircRNA loci
species–specific
shared
B: Dimer enrichment (shared vs. species-specific circRNA loci)
Mouse
Dimer count (species–specific circRNA loci)
Dimer count (shared circRNA loci)
0
10
20
B1_Mus1 B1_Mus2
B1_Mm B1_Mus1
B1_Mus1 B1_Mus1
B1_Mus2 B1_Mus2
B2_Mm2 B2_Mm2
p = 0.012
log2 enrichment
−2
0
−1
1
mean dimer age
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Rhesus
Dimer count (species–specific circRNA loci)
Dimer count (shared circRNA loci)
0
10
20
5
15
AluSp AluSx1
AluSz AluY
AluSx1 AluYra1
AluSx AluYra1
AluY AluYra1
p = 0.032
mean dimer age
1.0
1.5
2.0
2.5
3.0
3.5
4.0
log2 enrichment
−1.0
−0.5
0.0
0.5
1.0
C: Examples of repeat landscape
Human, AKT3 (chr1)
100 kb
243,550,000
243,600,000
243,650,000
243,700,000
243,750,000
243,800,000
243,850,000
Top-5 dimers
AluSx+AluSz
7x
AluSx1+AluSz
23x
AluSx+AluSx1
3x
AluSx+AluY
9x
AluSx1+AluY
18x
CircRNAs
hsCircRNA-1199
hsCircRNA-1198
hsCircRNA-3573
hsCircRNA-3574
RefSeq gene predictions (NCBI)
RefSeq Curated
Mouse, Akt3 (chr1)
100 kb
177,050,000
177,100,000
177,150,000
177,200,000
177,250,000
Top-5 dimers
B1_Mus2+B1_Mus2
2x
B1_Mus1+B1_Mus1
9x
B2_Mm2+B2_Mm2
1x
B1_Mus1+B1_Mus2
6x
CircRNAs
mmCircRNA-175
mmCircRNA-176
RefSeq gene predictions (NCBI)
UCSC RefSeq
Opossum, AKT3 (chr2)
100 kb
144,300,000
144,350,000
144,400,000
144,450,000
144,500,000
144,550,000
Top-5 dimers
SINE1_Mdo+SINE1_Mdo
50x
CircRNAs
mdCircRNA-213
mdCircRNA-212
mdCircRNA-700
Ensembl Gene Predictions - 95
ENSMODT00000008386.2

Figure 5. A: Degradation rates (MilliDivs) for top-5 dimers in rat and human. MilliDiv values for the top-5 dimers (defined by their presence in all parental genes) were compared between parental genes of species-specific (red) and shared (blue) circRNA loci in rat and human. Since dimers are composed of two repeats, their mean value was taken. A t-test was used to compare dimers between parental genes with shared and species-specific circRNA loci, with p-values plotted above the boxplots. Dimer order from left to right on the x-axis corresponds to their rank in the top-5 list (most frequent left) B: Dimer enrichment in shared vs. species-specific repeats in mouse and rhesus macaque. The frequency (number of detected dimers in a given parental gene), log2-enrichment (shared vs. species-specific) and mean age (defined as whether repeats are species-specific: age = 1, lineage-specific: age = 2, eutherian: age = 3, therian: age = 4) of the top-100 most frequent and least frequent dimers in parental genes with shared and species-specific circRNA loci in mouse and rhesus macaque were analysed. The frequency is plotted on the x- and y-axis, point size reflects the age and point colour the enrichment (blue = decrease, red = increase). Based on the comparison between shared and species-specific dimers, the top-5 dimers defined by frequency and enrichment are highlighted and labelled in red. C: Species-specific dimer landscape for the *Akt3* gene in human, mouse and opossum. UCSC genome browser view for the parental gene, circRNAs and top-5 dimers (as defined in panel B). Start and stop positions of each dimer are connected via an arc. Dimers are grouped by composition represented by different colours, the number of collapsed dimers is indicated to the right-side of the dimer group. Only dimers that start before and stop after a circRNAs are shown as these are potentially those that can contribute to the hairpin structure. The human *Akt3* gene possesses two circRNA clusters. For better visualisation, only the upstream cluster is shown.

Figure 5–Figure supplement 1. Species-specific repeats contribute to the formation of shared circRNA loci.

Discussion

Different scenarios have been proposed for how circRNA evolution takes place (see e.g. ***Patop et al.*** (***2019***) for a review). Our analyses of an extensive new cross-species dataset strongly suggest that many circRNA loci that are shared across orthologous genes of different species – of which there are surprisingly few – have emerged by convergent evolution, driven by structural commonalities of their parental genes, rather than having evolved from common ancestral circRNA loci. Parental genes are composed of many exons, are located in genomic regions of low GC content, and are surrounded by an elevated number of TEs, together creating "circRNA niches" – genomic regions in which circRNAs are more likely to be generated. TEs are an indispensable feature of the niche, and in addition to their similarity in structure, orthologous parental genes thus also possess a similar, pronounced integration bias for transposons, which subsequently manifests in genomic "TE hotspots" that are shared across species. Accordingly, many TEs found within the circRNA niche possess species-specific amplification patterns and have been active only recently, or are still active even today. Due to their evolutionary youth, the genomic sequences of TEs in the circRNA niche are barely degraded, increasing the likelihood of intramolecular RNA secondary structures, which have previously been associated with circRNA biogenesis. Taken together, these findings suggest that circRNAs and TEs co-evolve in a species-specific and dynamic manner. Moreover, as most circRNAs are evolutionarily young, they are overall rather unlikely to fulfil crucial functions. This idea is in agreement with the generally low expression levels of circRNAs and with accumulation patterns that are frequently tissue-specific and confined to post-mitotic cells (***Guo et al., 2014***; ***Westholm et al., 2014***). The model we present provides an explanation for how circRNAs can arise from shared (orthologous) exonic loci among species even if they themselves are not homologous (i.e., they do not stem from common evolutionary precursors that emerged in common ancestors). Finally, the properties we identified for the orthologous genomic niche can serve to predict circRNA parental genes with high confidence, opening the possibility to improve current circRNA prediction tools and to prioritise circRNAs for potential functional experiments.

TEs are a major component of most genomes and associated with various mechanisms that shape genome architecture and evolution. For example, TE integration into exons (changing the coding sequence) or at splice sites (potentially altering splicing patterns) may lead to the production of erroneous transcripts (***Zhang et al., 2011***). Other integration events are less sensitive towards creating such potentially hazardous “transcriptional noise”. For example, TEs that integrate in safe distance to important regions of a gene - e.g. in the middle of a long intron - might not cause more than a small increase in the transcript error rate that will in most cases be tolerable for the organism. As a consequence, TEs are more likely to be tolerated in genes with long introns than in short and compact genes. Moreover, long genes are known to be GC-poor (***Zhu et al., 2009***). These characteristics overlap precisely with those that we identify for circRNAs, which are also frequently generated from genes that are poor in GC and that have long introns, complex gene structures, as well as many TEs. In other words, the propensity to produce circRNAs scales with the same features that also predispose genes to transcriptional noise. Conceivably, many circRNAs may thus represent, at their core, a side effect of the genes’ transcriptional noise. In agreement with this model, a recent study in rat neurons has reported that the set of circRNAs that is upregulated after spliceosome inhibition is characterised by even longer flanking introns and an even higher number of RVCs than the average circRNA (***Wang et al., 2019***). Why is it frequently the same (orthologous) genomic loci and exons across species that independently develop the capacity for circRNA production? It is plausible that this phenomenon can be put down to tolerance for error rates. Let us consider repeat integration in close proximity to an exon boundary, which is an event that will likely alter local GC content. For example, GC-rich SINE elements that integrate in close proximity to a splice site can lead to a local increase in GC, which decreases the GC amplitude at the exon-intron boundary. Especially in GC-low genes, this can interfere with the intron-defined mechanism of splicing and cause mis-splicing (***Amit et al., 2012***). It is thus likely that TE integration close to a very strong splice site (i.e., with strong GC amplitude, as typically found in canonical exons) would have fewer repercussions on transcript error rates than integration close to alternative exons, whose GC amplitudes are less pronounced. Fully in line with such a model, we found that exons that are used in circRNAs are typically canonical exons with strong GC amplitudes. While at first sight, circRNA exons therefore appear to combine many rather specific, evolutionarily relevant properties (in particular, increased phastCons scores), we deem it probable that these are a mere consequence of a higher tolerance of canonical exon-flanking introns to TE integration.

Notably, this model may be taken even one step further by speculating whether circRNA properties for which a connection to TEs appears far-fetched, could in fact be ascribed to a transposon effect after all. Such cases are, for example, the reported predisposition of circRNAs to RNA editing (***Ivanov et al., 2015***) and different methylation patterns at both the RNA and DNA level (***Zhou et al., 2017***; ***Enuka et al., 2016***; ***Deniz et al., 2019***; ***Aktaş et al., 2017***). How could transposons come into play? Briefly, intronic TEs can facilitate the formation of local secondary structures in pre-mRNAs. On the one hand, this would interfere with splice-site accessibility and lead to an increase in transcript error rates (***Salari et al., 2012***; ***Melamud and Moult, 2009***). On the other hand, the secondary structures are associated with circRNA production. To avoid the negative impact of TEs on gene transcription, several defence mechanisms have evolved to silence them. RNA editing, for example, is thought to have evolved as a mechanism to suppress TE amplification, and A-to-I RNA editing is indeed associated with intronic Alu elements to inhibit Alu dimers (***Lev-Maor et al., 2008***; ***Athanasiadis et al., 2004***). In agreement with this notion, circRNA flanking introns are enriched in A-to-I editing sites, and knockdown of the editing machinery leads to an increase in circRNA levels (***Ivanov et al., 2015***; ***Rybak-Wolf et al., 2015***). Based on such findings, the conclusion has been drawn that A-to-I editing could represent a mechanism to control circRNA production (***Ivanov et al., 2015***; ***Rybak-Wolf et al., 2015***). However, the alternative scenario appears equally likely, in that changes in circRNA frequencies are actually a secondary effect caused by the primary purpose of A-to-I editing, namely the inhibition of Alu amplification. This notion is in line with the findings of (***Aktaş et al., 2017***) who showed that the nuclear RNA helicase DHX9 interacts with ADAR and can bind to inverted Alu elements that are transcribed as part of the gene (***Aktaş et al., 2017***). The loss of DHX9 leads to an increase of circRNA abundance from parental genes, in agreement with the model that DHX9 resolves TE-induced mRNA secondary structures to avoid interference with post-transcriptional processes (***Aktaş et al., 2017***). Similar reasoning can be applied to other modifications at the DNA and RNA level. Notably, DNA methylation interferes with TE amplification (***Yoder et al., 1997***), and has been connected to circRNA production (***Enuka et al., 2016***).

The modification N^6-methyladenosine (m^{6}A) plays various roles in mRNA metabolism, including in mRNA splicing, degradation and translation (reviewed in ***Zaccara et al.*** (***2019***)). m^{6}A is enriched in circRNA exons and can trigger circRNA cleavage and degradation (***Zhou et al., 2017***; ***Park et al., 2019***; ***Di Timoteo et al., 2020***) and has therefore been viewed as a way to control circRNA levels dynamically and in a tissue-specific manner. However, increased levels of m^{6}A, which is deposited already on the nascent RNA, are part of a much broader mechanism for mRNA destabilisation (reviewed in ***Lee et al.*** (***2020***)). Hence, it is possible that increased levels of m^{6}A on circRNAs rather reflect the general targeting of faulty transcripts for rapid degradation.

These considerations - together with our evolutionary data - lead us to the interpretation that many circRNAs likely represent transcriptional noise caused by TEs integrated into parental genes. However, it is also clear that molecular functions have been identified for several circRNAs (e.g. ***Hansen et al.*** (***2013***); ***Conn et al.*** (***2015***); ***Du et al.*** (***2016***)), although the absolute number of validated examples remains modest when compared to the high number of different circRNAs that have been detected across cell types, developmental stages and species. One would imagine that in order to evolve a function from noise, circRNAs need to reach critical, stable expression levels that bestow a positive effect on the organisms' fitness – a process that might take considerable time. Yet, circRNAs are not produced from scratch, but evolve from already existing functional genes, a process commonly known as exaptation (***Brosius and Gould, 1992***). A well-known example for this mechanism is provided by several miRNAs that evolved independently from each other in the same genomic position relative to the *Hox8* gene (***Campo-Paysaa et al., 2011***). For the circRNAs, the evolution of a function may be accelerated due to the presence of a clear exon structure and of regulatory elements from which the circRNAs can benefit. The production of structurally similar circRNAs from circRNA hotspots may accelerate this process, by providing different (back)splice sites and regulatory elements as evolutionary raw material, while keeping the internal exon sequence fairly similar. Once a circRNA emerges that is endowed with beneficial characteristics and equipped with an initial set of regulatory elements, the typically rather low expression level may increase. Robust expression and the acquisition of additional regulatory motifs (including those for RNA-binding proteins) may ultimately render the circRNA independent of its original regulation through reverse-complementary sequences (as described in ***Ashwal-Fluss et al.*** (***2014***); ***Conn et al.*** (***2015***); ***Okholm et al.*** (***2020***)). Thus, given that circRNAs are produced from hundreds of loci, in many cell types and across different developmental stages, beneficial circRNAs with useful functions – such as those that have been reported – may emerge and be fixed in a species during evolution.

In summary, our data suggests that many circRNA molecules do not carry specific molecular functions. However, one may still speculate whether it is actually the process of RNA circularization in itself, rather than the circRNA molecule, that is beneficial. For example, circularization may represent a mechanism to keep genes under control that have transformed into "transposon sinks", by directing mRNA output from such transposon-rich loci towards non-productive, circular transcripts. One could also argue that some level of splicing noise may be beneficial to engender gene expression plasticity at circRNA loci. Finally, circRNAs have emerged as reliable disease biomarkers (***Memczak et al., 2015***; ***Bahn et al., 2015***), and their utility for such predictive purposes is not affected by our conclusions – on the contrary. While an altered circRNA profile will likely not have a causal involvement in a disease, it could hint at misregulated transcription or splicing of the parental gene, at a novel TE integration event, or at problems with the RNA editing or methylation machinery. The careful analysis of the circRNA landscape may thus teach us about factors contributing to diseases in a causal fashion even if many or perhaps most circRNAs may not be functional but rather represent transcriptional noise.

Material and Methods

Data deposition, programmes and working environment

Table 2. Overview of external programmes.

Programme	Version
Blast	2.2.29+
BEDTools	2.17.0
Bowtie2	2.1.0
Clustal Omega	1.2.4
Cufflinks	2.1.1
FastQC	0.10.1
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

The raw data and processed data files discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (***Edgar et al., 2002***) and are accessible through the GEO Series accession number GSE162152. All scripts used to produce the main figures and tables of this publication have been deposited in the Git Repository circRNA_paperScripts. This Git repository also holds information on how to run the scripts, and links to the underlying data files for the main figures. The custom pipeline developed for the circRNA identification can be found in the Git Repository ncSplice_circRNAdetection.

Library preparation and sequencing

We used 5 μg of RNA per sample as starting material for library preparation, which were treated with 20 U RNase R (Epicentre/Illumina, Cat. No. RNR07250) for 1 h at 37°C to degrade linear RNAs, followed by RNA purification with the RNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's protocol. Paired-end sequencing libraries were prepared from the purified RNA with the Illumina TruSeq Stranded Total RNA kit with Ribo-Zero Gold according to the protocol with the following modifications to select larger 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 analysed on the fragment analyzer for their quality and sequenced with the Illumina HiSeq 2500 platform (multiplexed, 100 cycles, paired-end, read length 100 nt).

Identification and quantification of circRNAs

Mapping of RNA-seq data

The ensembl annotations for opossum (monDom5), mouse (mm10), rat (rn5), rhesus macaque (rheMac2) and human (hg38) were downloaded from Ensembl 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 3. 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

Analysis of unmapped reads

We developed a custom pipeline to detect circRNAs (**Figure1-Figure supplement 1B**), 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. For paired-end sequencing reads, the mate read not mapping to the backsplice junction can often be mapped to the reference genome without any problem. However, it will be classified as "unmapped read" (because its mate read mapping to the backsplice junction was not identified by the standard procedure). Next, all unpaired reads are thus selected from the accepted_hits.bam file generated by TopHat2 (singletons) and assessed for whether the mate read (second read of the 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 built 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 deposited in the Git Repository nc-

Splice_circRNAdetection.

Trimming of overlapping reads

Due to 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 our custom pipeline described above, and the step will be performed automatically if the scripts are run.

Calculation of CPM value

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

$$\begin{aligned}
counts &= \frac{counts\_rep1 + counts\_rep2 + counts\_rep3}{3} \\
totalMappedReads &= \frac{mappedReads\_rep1 + mappedReads\_rep2 + mappedReads\_rep3}{3} \\
CPM &= \frac{counts \cdot 10^6}{totalMappedReads}
\end{aligned}$$

Filtering of candidates based on CPM enrichment

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

$$enrichment = \frac{CPM\_RNaseR}{CPM\_untreated}$$

Candidates with a log2-enrichment of smaller 1.5, as well as less than 0.05 CPM, were removed.

Manual filtering steps

We observed several genomic loci in rhesus macaque and human that were highly enriched in reads for putative BSJs (no such problem was detected for opossum, mouse and rat). 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. These candidates were thus removed manually; the regions are:
All following analyses were conducted with the circRNA candidates that remained after this step.

Reconstruction of circRNA isoforms

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

Table 4. Removed regions during mapping.

species	tissue	chromosome	start	stop	strand
rhesus macaque	testis	7	164261343	164283671	+
rhesus macaque	testis	7	22010814	22092409	-
rhesus macaque	testis	19	52240850	52288425	-
rhesus macaque	testis	19	59790996	59834798	+
rhesus macaque	testis	19	59790996	59847609	+
human	testis	2	178535731	178600667	+
human	testis	7	66429678	66490107	-
human	testis	9	97185441	97211487	-
human	testis	12	97492460	97561047	+
human	testis	14	100913431	100949596	+
human	testis	18	21765771	21849388	+

$$geometric\_mean = \left(\prod x\right)^{\frac{1}{length(x)}}$$

$$size\_factor = median\left(\frac{x}{geometric\_mean}\right)$$

with *x* being a vector containing the number of reads per junction. We then compared read coverage for junctions outside and inside the BSJ for each gene and 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. The observed coverage change of junctions inside the backsplice was then compared to the expected change in the background distribution and junctions with a log2-change outside the 90% confidence interval were assigned as circRNA junctions; a loose cut-off was chosen, 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, we reconstructed a splicing graph for each circRNA candidate, in which network nodes are exons connected by splice junctions (edges) (***Heber et al., 2002***). 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. We used a simple breadth-first-search algorithm to traverse the graph and to define the strength for each possible isoform by its mean coverage. Only the strongest isoform was considered for all subsequent analyses.

Reconstruction and expression quantification of linear mRNAs

We reconstructed linear isoforms based on the pipeline provided by ***Trapnell et al.*** (***2012***) (Cufflinks + Cuffcompare + Cuffnorm). 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 normalised across species and tissues using a median scaling approach as described in ***Brawand et al.*** (***2011***).

Identification of shared circRNA loci between species

Definition and identification of shared circRNA loci

Shared circRNA loci 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 (see **Figure 2A** and **Figure 2-Figure supplement 1**). Overall considerations of this kind have recently also been outlined in ***Patop et al.*** (***2019***).

Level 1 - Parental genes: 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 retrieve the 1:1 orthologous genes for the eutherians (mouse, rat, rhesus macaque, human), for rodents (mouse, rat) and primates (rhesus macaque, human). Shared circRNA loci between species were assessed by counting the number of 1:1 orthologous parental genes between the five species. The analysis was restricted to protein-coding genes.

Level 2 - circRNA locus: 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, we 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*). Collapsed transcripts were defined as overlapping between different species if they shared at least one exon, independent of the exon length.

Level 3 - start/stop exon: To identify circRNAs sharing the same first and last exon between species, we lifted exons coordinates between species (same settings as described above, *liftOver, -minMatch=0.5*). The circRNA was then defined as "shared", if both exons were annotated as start and stop exons in the respective circRNAs of the given species. Note, that this definition only requires an overlap for start and stop exons, internal circRNA exons may differ.

Given that only circRNAs that comprise corresponding (1:1 orthologous exons) in different species might at least potentially and reasonably considered to be homologous (i.e., might have originated from evolutionary precursors in common ancestors) and the Level 3 definition might require strong evolutionary conservation of splice sites (i.e., with this stringent definition many shared loci may be missed), we decided to use the level 2 definition (circRNA locus) for the analyses presented in the main text, while we still provide the results for the Level 1 and 3 definitions in the supplement (**Figure 2-Figure supplement 1**). Importantly, defining shared circRNA loci at this level allows us to also compare circRNA hostspots which have been defined using a similar classification strategy.

Clustering of circRNA loci between species

Based on the species set in which shared circRNA loci were found, we categorised circRNAs in the following groups: Species-specific, rodent, primate, eutherian and therian circRNAs. To be part of the rodent or primate group, the circRNA has to be expressed in both species of the lineage. To be part of the eutherian group, the circRNA has to be expressed in three species out of the four species mouse, rat, rhesus macaque and human. To be part of the therian group, 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, we used the cluster algorithm MCL (***Enright et al., 2002***; ***Dongen, 2000***). MCL is frequently used to reconstruct orthology clusters based on blast results. It requires input in *abc* format (file: *species.abc*), 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 this analysis, 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
```

PhastCons scores

Codings exons were selected based on the attribute "transcript_biotype = protein_coding" in the gtf annotation file of the respective species and labelled as circRNA exons if they were in our circRNA annotation. Exons were further classified into UTR-exons and non-UTR exons using the ensembl field "feature = exon" or "feature = UTR". Since conservation scores are generally lower for UTR-exons (***Pollard et al., 2010***), any exon labelled as UTR-exon was removed from further analyses to avoid bias when comparing circRNA and non-circRNA exons. Genomic coordinates of the remaining exons were collapsed using the *merge* command from the BEDtools toolset (*bedtools merge input_file -nms -scores collapse*) to obtain a list of unique genomic loci. PhastCons scores for all exon types 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). For each gene type (parental or non-parental), the median phastCons score was calculated for each exon type within the gene (if non-parental: median of all exons; if parental: median of exons contained in the circRNA and median of exons outside of the circRNA).

Tissue specificity of exon types

Using the DEXseq package (from HTSeq 0.6.1), reads mapping on coding exons of the parental genes were counted. The exon-bins defined by DEXseq (filtered for bins >=10 nt) were then mapped and translated onto the different exon types: UTR-exons of parental genes, exons of parental genes that are not in a circRNA, circRNA exons. For each exon type, an FPKM value based on the exon length and sequencing depth of the library was calculated.

$$FPKM = \frac{counts\_for\_exon\_type \cdot 10^9}{exon\_type\_length/sequencing\_depth}$$

Exons were labelled as expressed in a tissue, if the calculated FPKM was at least 1. The maximum number of tissues in which each exon occurred was plotted separately for UTR-exons, exons outside the circRNA and contained in it.

GC amplitude

The ensembl annotation for each species was used to retrieve 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 (several values for the last *n* intronic bp and the first *m* exonic bp were tested beforehand, the 250:50 ratio was chosen, because it gave the strongest signal). 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.

Parental gene analysis

GC content of exons and intron

The ensembl annotation for each species was used to retrieve 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 further analyses. Differences between the GC content were assessed with a one-tailed Mann-Whitney U test.

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 the summed length of all alignments in a gene divided by its length (first to last exon).

Generalised linear models

All linear models were developed in the R environment. The presence of multicollinearity between predictors was assessed using the *vif()* function from the R package *car* (version 3.0-3) to calculate the variance inflation factor (VIF). Predictors were scaled to be able to compare them with each other using the *scale()* function as provided in the R environment.

For parental genes, the dataset was split into training (80%) and validation set (20%). To find the strongest predictors, we used the R package *bestglm* (version 0.37). Each model was fitted on the complete dataset using the command *bestglm()* with the information criteria set to “CV” (CV = cross validation) and the number of repetitions *t = 1000*. The model family was set to “binomial” as we were merely interested in predicting the presence (1) or absence (0) of a parental gene. Significant predictors were then used to report log-odds ratios and significance levels for the validation set using the default *glm()* function of the R environment. Log-odds ratios, standard errors and confidence intervals were standardised using the *beta()* function from the *reghelper* R package (version 1.0.0) and are reported together with their p-values in **Supplementary Table 5**.

For the correlation of hotspot presence across the number of species, a generalised linear model was applied using the categorical predictors “lineage” (= circRNA loci shared within rodents or primates), “eutherian” (= circRNA loci shared within rodents and primates) and “therian” (= circRNA loci shared within opossum, rodents and primates). Log-odds ratios, standard errors and confidence intervals were standardised using the *beta()* function from the *reghelper* R package (version 1.0.0) and are reported together with their p-values in **Supplementary Table 6**.

Comparison to human and mouse circRNA heart dataset

The circRNA annotations for human and mouse heart as provided by ***Werfel et al.*** (***2016***) were, based on the parental gene ID, merged with our circRNA annotations. Prediction values for parental genes were calculated using the same general linear regression models as described above (Section *Generalised linear models* in Material and Methods section) with genomic length, number of exons, GC content, expression levels, reverse complements (RVCs) and phastCons scores as predictors. Prediction values were received from the model and compared between parental genes predicted by our and the Werfel dataset as well as between the predictors in non-parental and parental genes of the Werfel dataset (**Figure 3-Figure supplement 3**).

Integration of external studies

(1) Replication time

Values for the replication time were used as provided in ***Koren et al.*** (***2012***). Coordinates of the different replication domains were intersected with the coordinates of coding genes using BEDtools (*bedtools merge -f 1*). The mean replication time of each gene was used for subsequent analyses.

(2) Gene expression steady-state levels

Gene expression steady-state levels and decay rates were used as provided in Table S1 of ***Pai et al.*** (***2012***).

(3) GHIS

Genome-wide haploinsufficiency scores for each gene were used as provided in Supplementary Table S2 of ***Steinberg et al.*** (***2015***).

Repeat analyses

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.

Repeat definition

The RepeatMasker annotation for full and nested repeats were downloaded for all genomes using the UCSC Table browser (tracks "RepeatMasker" and "Interrupted Rpts") and the two files merged. Nested repeats were included, because it was shown that small repetitive regions are sufficient to trigger base pairing necessary for backsplicing (***Liang and Wilusz, 2014***; ***Kramer et al., 2015***). The complete list was then intersected (*bedtools merge -f1*) with the above defined list of background and flanking introns for further analyses.

Identification of repeat dimers

The complementary regions (RVCs) that were defined with megaBLAST as described above, 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*). As RVCs can contain several repeats, the "strongest" dimer was selected based on the number of overlapping base pairs (= longest overlapping dimer). The "dimer list" obtained from this analysis for each species was further ranked according to the absolute frequency of each dimer. The proportion of the top-5 dimer frequency to all detected dimers, was calculated based on this list ($n_{top\text{-}5}$ / n_{all_dimers}).

Binding scores of repeat dimers

Binding scores for each TE class (based on the TE reference sequence) were defined by taking into account the (1) phylogenetic distance to other repeat families in the same species and (2) its binding affinity (deltaG) to those repeats. We decided to not include the absolute TE frequency into the binding score, because it is a function of the TE's age, its amplification and degradation rates. Simulating the interplay between these three components is not in scope of this study, and the integration of frequency into binding score creates more noise as tested via PCA analyses (variance explained drops by 10%).

(1) Phylogenetic distance

TE reference sequences were obtained from Repbase (***Bao et al., 2015***) and translated into fasta-format for alignment (*reference_sequences.fa*). Alignments were then generated with Clustal Omega (v1.2.4) (***Sievers et al., 2011***) using the following settings:

```
$ clustalo -i reference_sequences.fa –distmat-out = repeats.mat –guidetree-out = repeats.dnd –full
```

The resulting distance matrix for the alignment was used for the calculation of the binding score. Visualisation of the distance matrix (**Figure 4C**, **Figure 4-Figure supplement 1**) was performed using the standard R functions *dist(method="euclidian")* and *hclust(method="ward.D2")*. Since several TE classes evolved independently from each other, the plot was manually modified to remove connections or to add additional information on the TE's origin from literature.

(2) Binding affinity

To estimate the binding affinity of individual TE dimers, the free energy of the secondary structure of the respective TE dimers was calculated with the RNAcofold function from the ViennaRNA Package:

```
$ RNAcofold -a -d2 < dimerSequence.fa
```

with *dimerSequence.fa* containing the two reference sequences of the TEs from which the dimer is composed. The resulting deltaG values were used to calculate the binding score.

(3) Final binding score

To generate the final binding score, values from the distance matrix and the binding affinity were standardised (separately from each other) to values between 0 and 1:

$$f(x) = \frac{x - min(v)}{max(v) - min(v)}$$

with *x* being the binding affinity/dimer frequency and *minv* and *maxv* the minimal and maximal observed value in the distribution. The standardised values for the binding affinity and dimer frequency were then summed up (= binding score) and classified by PCA using the R environment:

```
$ pca <- prcomp(score, center=TRUE, scale.=FALSE)
```

PC1 and PC2 were used for subsequent plotting with the absolute frequency of dimers represented by the size of the data points.

Calculation of dimer degradation

RepeatMasker annotations were downloaded from the UCSC Table browser for all genomes. The milliDiv values for each TE in a TE dimer were retrieved from this annotation for full and nested repeats. A representative milliDiv was formed using the mean of the two values. Dimers were then classified as species-specific or present in all species based on whether the circRNA parental gene produced species-specific or shared circRNA loci. Significance levels for milliDiv differences between the dimer classes were assessed with a simple Mann-Whitney U test (alternative set to “less”).

Supplementary Data

Supplementary Tables and Figures

Supplementary Tables and Figures are available as an attachmente to this document.

Supplementary Files

Supplementary File 1: CircRNA annotation file for opossum. A gtf-file with all circRNA transcripts including the transcript and exon coordinates.

Supplementary File 2: CircRNA annotation file for mouse. A gtf-file with all circRNA transcripts including the transcript and exon coordinates.

Supplementary File 3: CircRNA annotation file for rat. A gtf-file with all circRNA transcripts including the transcript and exon coordinates.

Supplementary File 4: CircRNA annotation file for rhesus macaque. A gtf-file with all circRNA transcripts including the transcript and exon coordinates.

Supplementary File 5: CircRNA annotation file for human. A gtf-file with all circRNA transcripts including the transcript and exon coordinates.

All gtf-files have been uploaded to the UCSC genome browser and can be viewed here:

Opossum: http://genome.ucsc.edu/s/Frenzchen/monDom5%20circRNA%20annotation

Mouse; http://genome.ucsc.edu/s/Frenzchen/mm10%20circRNA%20annotation

Rat: http://genome.ucsc.edu/s/Frenzchen/rn5%20circRNA%20annotation

Rhesus macaque: http://genome.ucsc.edu/s/Frenzchen/rheMac2%20circRNA%20annotation

Human: http://genome.ucsc.edu/s/Frenzchen/hg38%20circRNA%20annotation

Acknowledgments

We thank the Lausanne Genomics Technologies Facility for high throughput sequencing support; Jean Halbert, Delphine Valloton and Angelica Liechti for opossum, mouse and rat tissue dissection and RNA extractions; Philipp Khaitovich for providing human and rhesus macaque samples; Bulak Arpat, Thomas O. Auer and Romane Meurs for discussions on the manuscript; and Ioannis Xenarios for discussion and support with IT-infrastructure and data archiving. D.G. acknowledges funding by the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) RNA and Disease (141735) and individual grant 179190. P.J. was supported by Human Frontiers Science Program long-term fellowship LT000158/2013-L. F.G. was supported by the SIB PhD Fellowship granted by the SIB Swiss Institute of Bioinformatics and the Fondation Leenards.

Author contributions

Contributions to this publication are distributed as follows: Study design: F.G., D.G., H.K. and P.J.; Experimental work: F.G. and P.J.; Bioinformatics data analyses: F.G.; Paper manuscript and discussion: F.G., D.G. and H.K.

Competing interests

No competing interests.

References

Aktaş T, Avşar Ilık Maticzka D, Bhardwaj V, Pessoa Rodrigues C, Mittler G, Manke T, Backofen R, Akhtar A. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. Nature. 2017 apr; 544(7648):115–119. http://dx.doi.org/10.1038/nature21715, doi: 10.1038/nature21715.

Alhasan AA, Izuogu OG, Al-Balool HH, Steyn JS, Evans A, Colzani M, Ghevaert C, Mountford JC, Marenah L, Elliott DJ, Santibanez-Koref M, Jackson MS. Circular RNA enrichment in platelets is a signature of transcriptome degradation. Blood. 2015 dec; http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=26660425&retmode=ref&cmd=prlinks, doi: 10.1182/blood-2015-06-649434.

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, Pupko T, Ast G. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell reports. 2012 may; 1(5):543–556. http://dx.doi.org/10.1016/j.celrep.2012.03.013, doi: 10.1016/j.celrep.2012.03.013.

Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A. circRNA Biogenesis Competes with Pre-mRNA Splicing. Molecular cell. 2014; .

Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biology. 2004 dec; 2(12):e391. http://dx.doi.org/10.1371/journal.pbio.0020391, doi: 10.1371/journal.pbio.0020391.

Bachmayr-Heyda A, Reiner AT, Auer K, Sukhbaatar N, Aust S, Bachleitner-Hofmann T, Mesteri I, Grunt TW, Zeillinger R, Pils D. Correlation of circular RNA abundance with proliferation–exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. Scientific reports. 2015; .

Bahn JH, Zhang Q, Li F, Chan TM, Lin X, Kim Y, Wong DTW, Xiao X. The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. Clinical Chemistry. 2015 jan; 61(1):221–230. http://dx.doi.org/10.1373/clinchem.2014.230433, doi: 10.1373/clinchem.2014.230433.

Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015 jun; 6:11. http://dx.doi.org/10.1186/s13100-015-0041-9, doi: 10.1186/s13100-015-0041-9.

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nature Reviews Genetics. 2002 may; 3(5):370–379. http://dx.doi.org/10.1038/nrg798, doi: 10.1038/nrg798.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H. The evolution of gene expression levels in mammalian organs. Nature. 2011 oct; 478(7369):343–348. http://dx.doi.org/10.1038/nature10532, doi: 10.1038/nature10532.

Brosius J, Gould SJ. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proceedings of the National Academy of Sciences of the United States of America. 1992 nov; 89(22):10706–10710. http://dx.doi.org/10.1073/pnas.89.22.10706, doi: 10.1073/pnas.89.22.10706.

Campo-Paysaa F, Sémon M, Cameron RA, Peterson KJ, Schubert M. microRNA complements in deuterostomes: origin and evolution of microRNAs. Evolution & Development. 2011 feb; 13(1):15–27. http://dx.doi.org/10.1111/j.1525-{142X}.2010.00452.x, doi: 10.1111/j.1525-{142X}.2010.00452.x.

Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, Roslan S, Schreiber AW, Gregory PA, Goodall GJ. The RNA binding protein quaking regulates formation of circRNAs. Cell. 2015 mar; 160(6):1125–1134. http://dx.doi.org/10.1016/j.cell.2015.02.014, doi: 10.1016/j.cell.2015.02.014.

Cortés-López M, Gruner MR, Cooper DA, Gruner HN, Voda AI, van der Linden AM, Miura P. Global accumulation of circRNAs during aging in Caenorhabditis elegans. BMC Genomics. 2018 jan; 19(1):8. http://dx.doi.org/10.1186/s12864-017-4386-y, doi: 10.1186/s12864-017-4386-y.

Deniz , Frost JM, Branco MR. Regulation of transposable elements by DNA modifications. Nature Reviews Genetics. 2019; 20(7):417–431. http://www.nature.com/articles/s41576-019-0106-6, doi: 10.1038/s41576-019-0106-6.

Di Timoteo G, Dattilo D, Centrón-Broco A, Colantoni A, Guarnacci M, Rossi F, Incarnato D, Oliviero S, Fatica A, Morlando M, Bozzoni I. Modulation of circRNA Metabolism by m(6)A Modification. Cell reports. 2020 May; 31(6):107641. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=32402287&retmode=ref&cmd=prlinks, doi: 10.1016/j.celrep.2020.107641.

Dongen S. Performance criteria for graph clustering and Markov cluster experiments. . 2000 May; http://dl.acm.org/citation.cfm?id=868979.

Du WW, Yang W, Liu E, Yang Z, Dhaliwal P, Yang BB. Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. Nucleic Acids Research. 2016 apr; 44(6):2846–2858. http://dx.doi.org/10.1093/nar/gkw027, doi: 10.1093/nar/gkw027.

Dubin RA, Kazmi MA, Ostrer H. Inverted repeats are necessary for circularization of the mouse testis Sry transcript. Gene. 1995 dec; 167(1-2):245–248. https://www.ncbi.nlm.nih.gov/pubmed/8566785.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002 jan; 30(1):207–210. http://dx.doi.org/10.1093/nar/30.1.207, doi: 10.1093/nar/30.1.207.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 2002 apr; 30(7):1575–1584. http://dx.doi.org/10.1093/nar/30.7.1575, doi: 10.1093/nar/30.7.1575.

Enuka Y, Lauriola M, Feldman ME, Sas-Chen A, Ulitsky I, Yarden Y. Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. Nucleic Acids Research. 2016 feb; 44(3):1370–1383. http://dx.doi.org/10.1093/nar/gkv1367, doi: 10.1093/nar/gkv1367.

Ermakova EO, Nurtdinov RN, Gelfand MS. Fast rate of evolution in alternatively spliced coding regions of mammalian genes. BMC Genomics. 2006 apr; 7:84. http://dx.doi.org/10.1186/1471-2164-7-84, doi: 10.1186/1471-2164-7-84.

Galtier N, Mouchiroud D. Isochore evolution in mammals: a human-like ancestral structure. Genetics. 1998 dec; 150(4):1577–1584. https://www.ncbi.nlm.nih.gov/pubmed/9832533.

Gruner H, Cortés-López M, Cooper DA, Bauer M, Miura P. CircRNA accumulation in the aging mouse brain. Scientific Reports. 2016 dec; 6:38907. http://dx.doi.org/10.1038/srep38907, doi: 10.1038/srep38907.

Gu W, Ray DA, Walker JA, Barnes EW, Gentles AJ, Samollow PB, Jurka J, Batzer MA, Pollock DD. SINEs, evolution and genome structure in the opossum. Gene. 2007 jul; 396(1):46–58. http://dx.doi.org/10.1016/j.gene.2007.02.028, doi: 10.1016/j.gene.2007.02.028.

Guo JU, Agarwal V, Guo H, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. Genome Biology. 2014 jul; 15(7):409. http://dx.doi.org/10.1186/s13059-014-0409-z, doi: 10.1186/s13059-014-0409-z.

Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. Natural RNA circles function as efficient microRNA sponges. Nature. 2013 mar; 495(7441):384–388. http://dx.doi.org/10.1038/nature11993, doi: 10.1038/nature11993.

Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA. Splicing graphs and EST assembly problem. Bioinformatics (Oxford, England). 2002; 18 Suppl 1:S181–8. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=12169546&retmode=ref&cmd=prlinks, doi: 10.1093/bioinformatics/18.suppl_1.s181.

Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C, Rajewsky N. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. Cell reports. 2015 jan; 10(2):170–177. http://dx.doi.org/10.1016/j.celrep.2014.12.019, doi: 10.1016/j.celrep.2014.12.019.

Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA (New York). 2013 feb; 19(2):141–157. http://dx.doi.org/10.1261/rna.035667.112, doi: 10.1261/rna.035667.112.

Kim J, Deininger PL. Recent amplification of rat ID sequences. Journal of Molecular Biology. 1996 aug; 261(3):322–327. http://dx.doi.org/10.1006/jmbi.1996.0464, doi: 10.1006/jmbi.1996.0464.

Kim J, Martignetti JA, Shen MR, Brosius J, Deininger P. Rodent BC1 RNA gene as a master gene for ID element amplification. Proceedings of the National Academy of Sciences of the United States of America. 1994 apr; 91(9):3607–3611. http://dx.doi.org/10.1073/pnas.91.9.3607, doi: 10.1073/pnas.91.9.3607.

Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. Differential relationship of DNA replication timing to different forms of human mutation and variation. American Journal of Human Genetics. 2012 dec; 91(6):1033–1040. http://dx.doi.org/10.1016/j.ajhg.2012.10.018, doi: 10.1016/j.ajhg.2012.10.018.

Kramer MC, Liang D, Tatomer DC, Gold B, March ZM, Cherry S, Wilusz JE. Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and SR proteins. Genes & Development. 2015 oct; 29(20):2168–2182. http://dx.doi.org/10.1101/gad.270421.115, doi: 10.1101/gad.270421.115.

Kristensen LS, Andersen MS, Stagsted LVW, Ebbesen KK, Hansen TB, Kjems J. The biogenesis, biology and characterization of circular RNAs. Nature Reviews Genetics. 2019 aug; 20(11):675–691. http://dx.doi.org/10.1038/s41576-019-0158-7, doi: 10.1038/s41576-019-0158-7.

Lee Y, Choe J, Park OH, Kim YK. Molecular Mechanisms Driving mRNA Degradation by m(6)A Modification. Trends in genetics : TIG. 2020 Mar; 36(3):177–188. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=31964509&retmode=ref&cmd=prlinks, doi: 10.1016/j.tig.2019.12.007.

Lev-Maor G, Ram O, Kim E, Sela N, Goren A, Levanon EY, Ast G. Intronic Alus influence alternative splicing. PLoS Genetics. 2008 sep; 4(9):e1000204. http://dx.doi.org/10.1371/journal.pgen.1000204, doi: 10.1371/journal.pgen.1000204.

Li S, Li X, Xue W, Zhang L, Yang LZ, Cao SM, Lei YN, Liu CX, Guo SK, Shan L, Wu M, Tao X, Zhang JL, Gao X, Zhang J, Wei J, Li J, Yang L, Chen LL. Screening for functional circular RNAs using the CRISPR–Cas13 system. Nature Methods. 2020; https://doi.org/10.1038/s41592-020-01011-4.

Liang D, Wilusz JE. Short intronic repeat sequences facilitate circular RNA production. Genes & Development. 2014 oct; 28(20):2233–2247. http://dx.doi.org/10.1101/gad.251926.114, doi: 10.1101/gad.251926.114.

Melamud E, Moult J. Structural implication of splicing stochastics. Nucleic Acids Research. 2009 aug; 37(14):4862–4872. http://dx.doi.org/10.1093/nar/gkp444, doi: 10.1093/nar/gkp444.

Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le Noble F, Rajewsky N. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013 mar; 495(7441):333–338. http://dx.doi.org/10.1038/nature11928, doi: 10.1038/nature11928.

Memczak S, Papavasileiou P, Peters O, Rajewsky N. Identification and characterization of circular rnas as a new class of putative biomarkers in human blood. Plos One. 2015 oct; 10(10):e0141214. http://dx.doi.org/10.1371/journal.pone.0141214, doi: 10.1371/journal.pone.0141214.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, Jurka J, Kamal M, Mauceli E, Searle SMJ, Sharpe T, Baker ML, Batzer MA, Benos PV, Belov K, Clamp M, et al. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 2007 may; 447(7141):167–177. http://dx.doi.org/10.1038/nature05805, doi: 10.1038/nature05805.

Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nature Genetics. 2003 jun; 34(2):177–180. http://dx.doi.org/10.1038/ng1159, doi: 10.1038/ng1159.

Okholm TLH, Sathe S, Park SS, Kamstrup AB, Rasmussen AM, Shankar A, Chua ZM, Fristrup N, Nielsen MM, Vang S, Dyrskjøt L, Aigner S, Damgaard CK, Yeo GW, Pedersen JS. Transcriptome-wide profiles of circular RNA and RNA-binding protein interactions reveal effects on circular RNA biogenesis and cancer pathway expression. Genome Medicine. 2020; 12(1):112. https://doi.org/10.1186/s13073-020-00812-8.

Pai AA, Cain CE, Mizrahi-Man O, De Leon S, Lewellen N, Veyrieras JB, Degner JF, Gaffney DJ, Pickrell JK, Stephens M, Pritchard JK, Gilad Y. The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. PLoS Genetics. 2012 oct; 8(10):e1003000. http://dx.doi.org/10.1371/journal.pgen.1003000, doi: 10.1371/journal.pgen.1003000.

Park OH, Ha H, Lee Y, Boo SH, Kwon DH, Song HK, Kim YK. Endoribonucleolytic Cleavage of m(6)A-Containing RNAs by RNase P/MRP Complex. Molecular cell. 2019 May; 74(3):494–507.e8. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=30930054&retmode=ref&cmd=prlinks, doi: 10.1016/j.molcel.2019.02.034.

Patop IL, Wüst S, Kadener S. Past, present, and future of circRNAs. The EMBO Journal. 2019; 38(16):e100836. https://www.embopress.org/doi/abs/10.15252/embj.2018100836, doi: https://doi.org/10.15252/embj.2018100836.

Piwecka M, Glažar P, Hernandez-Miranda LR, Memczak S, Wolf SA, Rybak-Wolf A, Filipchyk A, Klironomos F, Cerda Jara CA, Fenske P, Trimbuch T, Zywitza V, Plass M, Schreyer L, Ayoub S, Kocks C, Kühn R, Rosenmund C, Birchmeier C, Rajewsky N. Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. Science. 2017 sep; 357(6357). http://dx.doi.org/10.1126/science.aam8526, doi: 10.1126/science.aam8526.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research. 2010 jan; 20(1):110–121. http://dx.doi.org/10.1101/gr.097857.109, doi: 10.1101/gr.097857.109.

Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, Hanan M, Behm M, Bartok O, Ashwal-Fluss R, Herzog M, Schreyer L, Papavasileiou P, Ivanov A, Öhman M, Refojo D, Kadener S, Rajewsky N. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Molecular Cell. 2015 jun; 58(5):870–885. http://dx.doi.org/10.1016/j.molcel.2015.03.027, doi: 10.1016/j.molcel.2015.03.027.

Salari R, Wojtowicz D, Zheng J, Levens D, Pilpel Y, Przytycka TM. Teasing apart translational and transcriptional components of stochastic variations in eukaryotic gene expression. PLoS Computational Biology. 2012 aug; 8(8):e1002644. http://dx.doi.org/10.1371/journal.pcbi.1002644, doi: 10.1371/journal.pcbi.1002644.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research. 2005 aug; 15(8):1034–1050. http://dx.doi.org/10.1101/gr.3715005, doi: 10.1101/gr.3715005.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011 oct; 7:539. http://dx.doi.org/10.1038/msb.2011.75, doi: 10.1038/msb.2011.75.

Smit A, Hubley R, Green P. RepeatMasker Open-4.0, 2013-2015. . 2013; http://www.repeatmasker.org.

Starke S, Jost I, Rossbach O, Schneider T, Schreiner S, Hung LH, Bindereif A. Exon circularization requires canonical splice signals. Cell reports. 2015; .

Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic Acids Research. 2015 sep; 43(15):e101. http://dx.doi.org/10.1093/nar/gkv474, doi: 10.1093/nar/gkv474.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012 mar; 7(3):562–578. http://dx.doi.org/10.1038/nprot.2012.016, doi: 10.1038/nprot.2012.016.

VenøMT, Hansen TB, VenøST, Clausen BH, Grebing M, Finsen B, Holm IE, Kjems J. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. Genome Biology. 2015 nov; 16:245. http://dx.doi.org/10.1186/s13059-015-0801-3, doi: 10.1186/s13059-015-0801-3.

Wang M, Hou J, Müller-McNicoll M, Chen W, Schuman EM. Long and Repeat-Rich Intronic Sequences Favor Circular RNA Formation under Conditions of Reduced Spliceosome Activity. iScience. 2019 oct; 20:237–247. http://dx.doi.org/10.1016/j.isci.2019.08.058, doi: 10.1016/j.isci.2019.08.058.

Wang PL, Bao Y, Yee MC, Barrett SP, Hogan GJ, Olsen MN, Dinneny JR, Brown PO, Salzman J. Circular RNA is expressed across the eukaryotic tree of life. Plos One. 2014 mar; 9(6):e90859. http://dx.doi.org/10.1371/journal.pone.0090859, doi: 10.1371/journal.pone.0090859.

Werfel S, Nothjunge S, Schwarzmayr T, Strom TM, Meitinger T, Engelhardt S. Characterization of circular RNAs in human, mouse and rat hearts. Journal of Molecular and Cellular Cardiology. 2016 jul; 98:103–107. http://dx.doi.org/10.1016/j.yjmcc.2016.07.007, doi: 10.1016/j.yjmcc.2016.07.007.

Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, Lai EC. Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell reports. 2014 dec; 9(5):1966–1980. http://dx.doi.org/10.1016/j.celrep.2014.10.062, doi: 10.1016/j.celrep.2014.10.062.

Wilusz JE. Repetitive elements regulate circular RNA biogenesis. Mobile genetic elements. 2015 jun; 5(3):1–7. http://dx.doi.org/10.1080/{2159256X}.2015.1045682, doi: 10.1080/{2159256X}.2015.1045682.

Xu K, Chen D, Wang Z, Ma J, Zhou J, Chen N, Lv L, Zheng Y, Hu X, Zhang Y, Li J. Annotation and functional clustering of circRNA expression in rhesus macaque brain during aging. Cell discovery. 2018 sep; 4:48. http://www.nature.com/articles/s41421-018-0050-1, doi: 10.1038/s41421-018-0050-1.

Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. Trends in Genetics. 1997 aug; 13(8):335–340. http://dx.doi.org/10.1016/s0168-9525(97)01181-5, doi: 10.1016/s0168-9525(97)01181-5.

Yoshimoto R, Rahimi K, Hansen TB, Kjems J, Mayeda A. Biosynthesis of Circular RNA ciRS-7/CDR1as Is Mediated by Mammalian-wide Interspersed Repeats. iScience. 2020; 23(7):101345. http://www.sciencedirect.com/science/article/pii/S2589004220305320, doi: https://doi.org/10.1016/j.isci.2020.101345.

You X, Vlatkovic I, Babic A, Will T, Epstein I, Tushev G, Akbalik G, Wang M, Glock C, Quedenau C, Wang X, Hou J, Liu H, Sun W, Sambandan S, Chen T, Schuman EM, Chen W. Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. Nature Neuroscience. 2015 apr; 18(4):603–610. http://dx.doi.org/10.1038/nn.3975, doi: 10.1038/nn.3975.

Zaccara S, Ries RJ, Jaffrey SR. Reading, writing and erasing mRNA methylation. Nature reviews Molecular cell biology. 2019; 20(10):608–624. https://doi.org/10.1038/s41580-019-0168-5.

Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, Yang L. Complementary sequence-mediated exon circularization. Cell. 2014 sep; 159(1):134–147. http://dx.doi.org/10.1016/j.cell.2014.09.001, doi: 10.1016/j.cell.2014.09.001.

Zhang Y, Romanish MT, Mager DL. Distributions of transposable elements reveal hazardous zones in mammalian introns. PLoS Computational Biology. 2011 may; 7(5):e1002046. http://dx.doi.org/10.1371/journal.pcbi.1002046, doi: 10.1371/journal.pcbi.1002046.

Zhou C, Molinie B, Daneshvar K, Pondick JV, Wang J, Van Wittenberghe N, Xing Y, Giallourakis CC, Mullen AC. Genome-Wide Maps of m6A circRNAs Identify Widespread and Cell-Type-Specific Methylation Patterns that Are Distinct from mRNAs. Cell reports. 2017 aug; 20(9):2262–2276. http://dx.doi.org/10.1016/j.celrep.2017.08.027, doi: 10.1016/j.celrep.2017.08.027.

Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics. 2009 jan; 10:47. http://dx.doi.org/10.1186/1471-2164-10-47, doi: 10.1186/1471-2164-10-47.

Figure 1–Figure supplement 1.

Figure 1–Figure supplement 2.

Figure 1–Figure supplement 3.

Figure 1–Figure supplement 4.

Figure 2–Figure supplement 1.

Figure 2–Figure supplement 2.

Figure 3–Figure supplement 1.

Figure 3–Figure supplement 2.

Figure 3–Figure supplement 3.

Figure 4–Figure supplement 1.

Figure 4–Figure supplement 2.

Figure 5–Figure supplement 1.