Non-coding AUG circRNAs constitute an abundant and conserved subclass of circles.

Lotte VW Stagsted[1], Katrine M Nielsen[1], Iben Daugaard[1], and Thomas B Hansen[1*]

1. Department of Molecular Biology and Genetics (MBG), and Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Aarhus, Denmark

* To whom correspondence should be addressed:
T.B.H.: tbh@mbg.au.dk;

Running title: AUG circRNAs

Keywords: circRNA, circRNA translation, circRNA biogenesis, non-coding RNA, dhx9, ribosome profiling, riboseq

Abstract

Circular RNAs (circRNAs) are a subset of non-coding RNAs (ncRNAs) previously considered as products of missplicing. Now, circRNAs are considered functional molecules, although to date, only few functions have been experimentally validated, and therefore the vast majority of circRNAs are without known relevance. Here, based on RNA sequencing from the ENCODE consortium, we identify and characterize a subset of circRNAs, coined AUG circRNAs, encompassing the annotated translational start codon from the protein-coding host genes. AUG circRNAs are more abundantly expressed and conserved than other groups of circRNAs, and they display an *Alu*-independent mechanism of biogenesis. The AUG circRNAs contain part of bona fide ORF, and in the recent years, several studies have reported cases of circRNA translation. However, using thorough cross-species analysis, extensive ribosome profiling, proteomics analyses, and experimental data on a selected panel of AUG circRNAs, we observe no indications of translation of AUG circRNAs or any other circRNAs. Our data provide a comprehensive classification of circRNAs and, collectively, the data suggest that the AUG circRNAs constitute an abundant subclass of circRNAs produced independently of primate-specific *Alu* elements. Moreover, AUG circRNAs exhibit high cross-species conservation and are therefore likely to be functionally relevant.

Introduction

Non-coding RNAs (ncRNAs) constitute the vast majority of the human transcriptome as only a few percent of the produced transcripts are translated into proteins (ENCODE Project Consortium 2012)**.** NcRNAs represent a highly heterogeneous group of molecules that besides including essential elements of protein synthesis, ribosomal RNA (rRNA) and transfer RNA (tRNA), also comprise small RNAs, such as microRNAs (miRNAs), which are involved in regulation of mRNA stability and protein synthesis (Bartel 2009), as well as long non-coding RNAs (lncRNAs) with many established functionalities (Böhmdorfer and Wierzbicki 2015; Noh et al. 2018). Recently, by means of high throughput tools, circular RNAs (circRNAs) were added to the rapidly expanding list of non-coding RNAs (Ebbesen et al. 2016). CircRNAs are typically derived from annotated protein-coding genes, but due to their relatively low abundance compared to their linear messenger RNA (mRNA) counterparts, circRNA molecules were first presumed to be missplicing events of the spliceosome with little to no relevance (Cocquerelle et al. 1993; Zaphiropoulos 1997). While this may be the case

46 for a substantial subset of circRNAs, the identification and functional characterization of the highly
47 conserved circRNA and miR-7-sponge, CDR1as/ciRS-7 (Hansen et al. 2013b; Memczak et al. 2013),
48 as well as extensive profiling of differentially expressed circRNAs from RNA sequencing analyses
49 (Rybak-Wolf et al. 2015; Memczak et al. 2013; Salzman et al. 2012; Veno et al. 2015) strongly
50 supports circRNAs as biologically relevant RNA species in eukaryotic cells. CircRNAs are generated
51 by non-linear splicing (coined backsplicing) where an upstream splice acceptor (SA) is covalently
52 joined to a downstream splice donor (SD) resulting in a circular structure (Hansen et al. 2011; Jeck
53 et al. 2013). This results in a very high intracellular stability due to the lack of free ends, which
54 protects them from normal exonucleolytic decay. CircRNAs are mostly comprised of exonic regions
55 (most commonly 2-3 exons) derived from annotated protein-coding transcripts (Zhang et al. 2014).
56 The current model of biogenesis suggests that backsplicing is stimulated by bringing the involved
57 splice sites into close proximity (Ebbesen et al. 2016). This is conventionally facilitated by inverted
58 *Alu* elements (IAE) (Jeck et al. 2013; Zhang et al. 2014), however trans-acting RNA-binding factors
59 have also been implicated in circRNA formation (Ashwal-Fluss et al. 2014; Conn et al. 2015; Li et al.
60 2017).
61 With the exception of the exon-intron circRNAs (EIciRNA) (Li et al. 2015), circRNAs are exported to
62 the cytoplasm (Jeck et al. 2013) with a recent study reporting the *Drosophila* Hel25E and its human
63 homologs to regulate nuclear export of circRNAs in a length-dependent manner (Huang et al. 2018).
64 In the cytoplasm, circRNAs have been shown to tether and 'sponge' miRNAs, initially exemplified by
65 CDR1as/ciRS-7 harbouring >70 miR-7 binding sites (Hansen et al. 2013a; Memczak et al. 2013). Since
66 then, several other examples have been published showing anti-miR effects of circRNA expression
67 (Peng et al. 2016; Chaiteerakij et al. 2017; Zheng et al. 2016), although, bioinformatics analysis
68 indicates that - apart from ciRS-7 - miRNA binding sites are generally not enriched in circRNA more
69 than expected by chance (Guo et al. 2014). CircRNAs can also sequester RNA binding proteins and
70 hereby modulate protein activity (Ashwal-Fluss et al. 2014). Additionally, synthetic circRNAs have
71 been engineered to express protein by the use of internal ribosome entry sites (IRESs) allowing cap-
72 independent translation (Wang and Wang 2015). Recently, it was shown that open reading frames
73 (ORFs) within endogenously expressed circRNAs give rise to circRNA-specific peptides (Legnini et al.
74 2017; Pamudurti et al. 2017; Yang et al. 2017; Zhang et al. 2018a; Yang et al. 2018; Zhang et al.
75 2018b) suggesting that circRNAs are not necessarily exclusively non-coding.

76 In this study, publicly available RNA sequencing datasets from the ENCODE consortium are used to
77 characterize the circRNA transcriptomes in 378 human and 75 murine samples, and the most
78 abundant circRNAs in each dataset are identified, analyzed and stratified based on their genomic
79 features. These analyses reveal that a substantial fraction of highly abundant circRNAs derives from
80 exons encoding the translational start codon, here coined AUG circRNAs. In addition, the AUG
81 circRNAs are more conserved than other groups of circRNAs and generally rely on an IAE-
82 independent mode of biogenesis. Lastly, to determine the protein-codon ability of AUG circRNAs,
83 we conduct extensive analyses of cross-species conservation and ribosome profiling (RiboSeq). This
84 shows that ORF-associating features are not preserved in evolution and that backsplice-spanning
85 reads found in RiboSeq datasets are not derived from translating ribosomes. Consistently, we fail to
86 detect any peptides derived from AUG circRNAs using ectopic overexpression in cell-lines.
87 Collectively, these results suggest that circRNAs are generally not subjected to translation and thus
88 the functional relevance of the most conserved and abundant AUG circRNAs remains elusive.

89

90 **Results**

91 *The ENCODE circRNA landscape*

92 To obtain a comprehensive overview of circRNA expression across multiple tissues and cell lines, we
93 took advantage of the total RNA sequencing datasets on human and mouse samples made available
94 from the ENCODE consortium (see Supplementary Table 1). We conducted circRNA prediction and
95 quantification using two established pipelines; find_circ (Memczak et al. 2013) and circexplorer2
96 (Zhang et al. 2016a). In total, find_circ and circexplorer2 identify 140,304 and 235,179 unique
97 circRNAs using slightly modified settings (see methods), respectively, of which 81,589 are shared by
98 both algorithms (**Fig. 1A**). The notable fraction of circRNAs only predicted by one algorithm – the
99 socalled exotic circRNAs – is in general lowly expressed (**Fig. 1B**), which is also reflected by a small
100 subset of exotic circRNAs in the top 1000 expressed circRNA candidates predicted by each algorithm
101 (1-8%, data not shown). Consistently, we observe a high positive correlation between the algorithms
102 for the abundant circRNA species (**Fig. 1C**, similar analyses for mouse samples are shown in
103 **Supplementary Figs. 1A-C**). We have previously shown that exotic circRNAs are more likely to be

104 false positives (Hansen et al. 2015), and therefore we decided to focus only on the circRNAs jointly
105 predicted by both algorithms.

106 The used ENCODE data comprises 378 samples derived from 218 different human tissues and cell
107 lines (or 75 samples from 26 tissues in mouse, **Supplementary Table 1**). Plotting the expression of
108 circRNAs in each sample reveal a marked difference in circRNA expression between the samples
109 with particular abundant circRNA expression levels in tissues (**Supplementary Figs. 2A** and **C**),
110 whereas the similar analysis for mRNAs show similar expression in all biosample types
111 (**Supplementary Figs. 2B** and **D**). Moreover, even though the detected diversity of circRNAs is much
112 lower in mouse, the overall expression levels are comparable (**Supplementary Fig. 3A**). CircRNA
113 levels have previously been correlated with proliferation, i.e. circRNAs tend to accumulate in slow
114 or non-proliferative tissue (Bachmayr-Heyda et al. 2015). Thus, circRNA profiling from non-dividing
115 cells may dominate the average expression levels of circRNA across samples. Instead of comparing
116 expression across samples, we instead focused on the highest expressed circRNA in each sample
117 (the alpha circRNA). Here, the alpha circRNA in many samples exhibits disproportionally high
118 expression compared to the bulk of circRNAs. In fact, assuming a log-normal distribution of circRNA
119 expression, the alpha circRNAs are significant outliers in more than half of the samples (246 out of
120 378, fdr<0.05, one-tailed Grubbs test), whereas only 4 of 378 samples show similar significant outlier
121 mRNAs. This tendency is also observed in mouse (**Supplementary Fig. 3B**).

122 Based on the ENCODE data, circHIPK3 is the most predominant alpha circRNA followed by the miR-
123 7 sponge, ciRS-7 (**Fig. 1D**). Even though most of the top 10 alpha circRNAs are found in the mouse
124 dataset, only circSLC8A1 and circCDYL are shared in the top 10 between mouse and human (**Fig. 1D**
125 and **Supplementary Fig. 4A**).

126 We then zoomed in on the top 10 alpha circRNAs, i.e. the ten circRNAs most often seen as the
127 highest expressed in a given sample, to determine the genomic features associating with these
128 highly abundant circRNA species (see **Fig. 1E**). Here, the human alpha circRNAs are flanked with very
129 distal inverted *Alu* elements (IAE), which is in stark contrast to the bulk of circRNAs (**Fig. 1F**) and the
130 prevalent model of biogenesis (Jeck et al. 2013; Zhang et al. 2014). Moreover, no significant
131 association between circRNA producing loci and other inverted repeat elements are observed for
132 the alpha circRNAs specifically or circRNAs in general compared to host gene exons (**Fig. 1F**). In

133 mouse, no repetitive elements are selectively demarcating circRNAs from host exons
134 (**Supplementary Fig. 4B**), although a slight tendency towards proximal B1/alu SINE elements was
135 detected. Instead, for both species, we observe a clear tendency for alpha circRNAs to have very
136 long flanking introns (**Fig. 1F** and **Supplementary Fig. 4B**, ciRS-7, circRMST and circFAT3 are without
137 annotated flanking intron and thus excluded in this analysis). Moreover, a positive correlation
138 between intron length and IAE distance is detected (**Fig. 1G** and **Supplementary Fig. 4C**), indicating
139 that circRNAs either utilize an *Alu*-dependent mechanism of biogenesis or require long flanking
140 introns to favour back-splicing.

141

142 *AUG circRNAs are highly expressed and conserved*

143 The vast majority of circRNAs derive from annotated splice-sites (Zhang et al. 2016a), and we
144 decided to stratify circRNAs by host gene annotation (see **Fig. 2A**). Here, circRNAs derived from
145 exons containing annotated start codons, coined AUG circRNA, comprise 5 out of the top 10 alpha
146 circRNAs in both human and mouse samples, whereas the percentage of AUG circRNA in general is
147 7-11% (**Fig. 2B** and **Supplementary Fig. 5A**). In fact, the AUG circRNAs in both human and mouse
148 also show a significant over-representation in subsets of highly expressed circRNAs, which is not
149 seen for other circRNA sub-classes (**Fig. 2B** and **Supplementary Fig. 5A**). Consistently, in both human
150 and mouse AUG circRNAs are generally and significantly more abundant than other circRNAs
151 (**Supplementary Figs. 6A-B and D-E**) both in terms of absolute expression but also regarding the
152 circular-to-linear ratios, whereas the host genes, from which the AUG circRNAs are derived, are not
153 significantly more abundant than for other circRNAs (**Supplementary Figs. 6C and F).**

154 Using liftover (UCSC) we evaluated the number of human circRNAs re-identified in the mouse
155 dataset of circRNAs as a measure of conserved biogenesis. Similar to AUG circRNAs, the fraction of
156 conserved circRNAs increase with expression (**Fig. 2C**). Focusing specifically on the top 1000 most
157 abundant human circRNAs based on total backsplice junction (BSJ)-spanning reads, 39% of all AUG
158 circRNAs are conserved comprising almost twice as many conserved species compared to the other
159 circRNA sub-groups (**Fig. 2D**, p=9.4e-4, Fisher's exact test), despite the fact that the 5'UTR-
160 embedded splice acceptors in human AUG circRNAs are often (21%) not annotated as splice sites in
161 mouse. As with the top 10 alpha circRNAs, AUG circRNAs generally exhibit distal IAE and longer

162 flanking introns (**Fig. 2E-F**). In fact, AUG circRNAs, conserved circRNAs and host gene exons exhibit
163 an overall similar distribution of IAEs (**Fig. 2E**). In contrast, the flanking intron lengths effectively
164 demarcate AUG circRNA and conserved circRNAs from host gene exons (**Fig. 2F**), which is also
165 supported by the analysis of AUG circRNAs in mouse (**Supplementary Fig. 5B**). In fact, using an atlas
166 of circRNAs from *Drosophila* (Westholm et al. 2014), AUG circRNAs also increase in frequency with
167 circRNAs expression and associate with significantly longer introns compared to other circRNAs
168 (**Supplementary Figs. 5C-D**), whereas fruit fly circles in general show no evidence of flanking
169 inverted elements (Westholm et al. 2014). Thus, based on the observations that AUG circRNAs are
170 more conserved and overall devoid of flanking inverted *Alu* elements, and the fact that AUG circRNA
171 features may extent all the way to invertebrates, we propose that AUG circRNAs are more likely to
172 be biologically relevant and to utilize an *Alu*-independent biogenesis pathway.

173 To demarcate *Alu*-dependent from *Alu*-independent circRNAs, we empirically determined the
174 distance to nearest IAE by which the cumulative fraction of circularizing exons differed the most
175 from non-circularizing host exons. Here, approximately 45% of all human circRNAs has an inverted
176 *Alu* element within 2300 nucleotides total distance, whereas this cut-off only applies to 22% of host
177 exons (**Supplementary Fig. 7A**). We thus defined this 45% subset as the *Alu*-dependent circRNAs.
178 Based on this demarcation, only 13% of *Alu*-dependent circRNAs are observed in mouse, which is
179 consistent with the fact that *Alu* elements are primate specific (**Supplementary Fig. 7C**). In contrast,
180 41% of *Alu*-independent and 31% of circRNAs with long flanking introns (defined empirically as
181 flanking intron > 6500 nts, **Supplementary Fig. 7B**) are conserved (**Supplementary Fig. 7B**), which
182 suggests that at least for the evolutionary relevant circRNAs, biogenesis relies more on having long
183 flanking introns instead of proximal inverted *Alu* repeats. However, the requirement for long
184 flanking introns in circRNA biogenesis is currently unclear, and therefore the mechanism governing
185 production of the majority of abundant and conserved circRNAs remains undisclosed.

186 Recently, the RNA resolvase, Dhx9, was shown to inhibit circRNA production by unwinding and
187 destabilizing RNA structures formed by inverted *Alu* elements in flanking regions of circRNAs (Aktaş
188 et al. 2017). Dhx9 is proposed to protect cells from adverse secondary structures in the nucleus. As
189 a consequence, circRNAs sensitive to Dhx9 depletion are considered products of aberrant
190 backsplicing mediated by random insertion of inverted repeat elements, and therefore these

191 circRNAs are more likely to be functionally irrelevant. Based on RNAseq from Dhx9-depleted HEK293
192 cells (**Supplementary Table 3**), it is possible to determine the subset of circRNAs sensitive to Dhx9
193 expression. We selected the top 1000 expressed circRNAs from this experiment (**Supplementary**
194 **Table 4**). Here, roughly 25% (275 circRNAs out of 1000) responds significantly (fdr < 0.05) to the
195 Dhx9 depletion (**Fig. 3A**), with a clear tendency towards proximal IAE and short flanking introns (**Figs.**
196 **3B-C**). Consistently, 39% of circRNAs designated as *Alu* dependent (IAE distance < 2300nt) are Dhx9-
197 sensitive compared to only 12% of the non-*Alu* circRNAs, whereas long flanking introns are generally
198 insensitive to Dhx9 compared to short introns (43% vs 23%, **Supplementary Fig. 8D**). Interestingly,
199 in alignment with the analyses described above, AUG circRNAs are significantly reduced in the Dhx9-
200 sensitive fraction (5% vs 15%, p=4.8e7, **Fig. 3D**). In fact, only 10% of the AUG circRNAs compared to
201 26% of non-AUG circRNAs is affected significantly in expression upon Dhx9 depletion (**Fig. 3E**).

202 CircHIPK3 has previously been characterized as an *Alu*-dependent circRNA (Zheng et al., 2016),
203 however, it is also an AUG circRNA, as well as the overall highest expressed circRNA in the ENCODE
204 data. As such, the most proximal IAEs are within 2300 nt (see **Fig. 3B**), and while these *Alu* elements
205 could stimulate biogenesis, circHIPK3 carries more than 60 *Alu* elements in the immediate flanking
206 introns and is insensitive to Dhx9 depletion (**Fig. 3A**). Dhx9-insensitive biogenesis is also observed
207 for the two additional top 10 expressed AUG circRNAs, circSETD3 and circVRK1. Instead, the three
208 AUG circRNAs in the top 10 fraction all associate with very long flanking introns (**Fig. 3B**), and
209 notably, circZBTB44, which in this analysis is termed 'ambiguous' because it overlaps both an AUG
210 and a non-coding transcript, share the same features. Consistently, HITS-CLIP analysis of Dhx9
211 occupancy shows a clear selection for binding in the immediate flanking regions of *Alu*-dependent
212 compared to *Alu*-independent circRNAs (**Fig. 3F** and **Supplementary Fig. 8E**), but also a clear
213 preference for non-AUG over AUG circRNAs (**Fig. 3F** and **Supplementary Fig. 8F**). Collectively, this
214 strongly indicates that the AUG circRNAs are generally not affected by Dhx9 helicase activity and
215 thus not depending on IAE for biogenesis.

216

217 *No detectable protein production from ectopically expressed AUG circRNAs*

218 Recent studies have shown that circRNAs despite lacking a 5'cap and 3'poly(A)-tail are still capable
219 of recruiting ribosomes and act as templates for protein synthesis (Legnini et al. 2017; Yang et al.

220 2017; Pamudurti et al. 2017). This is most likely facilitated by IRES-like elements in the circRNA
221 required for cap-independent translation. The AUG circRNAs all contain a 5' part of a bona fide ORF,
222 and translation of this putative ORF will in most cases produce a truncated protein mimicking the
223 N-terminal part of the host gene encoded protein. To test the hypothesis that AUG circRNAs are in
224 fact protein-coding circRNAs, we initially focused on the two-exon AUG circRNAs derived from the
225 LPAR1 gene (**Fig. 4A**). In the ENCODE data, circLPAR1 is the highest expressed circRNA in 21 samples,
226 and it was the most abundant circRNA in one of the first global analyses of circRNA expression in a
227 human fibroblast cell line, hs68 (Jeck et al. 2013). Here, circLPAR1 was shown to be 3-fold higher
228 expressed than the second highest circRNA and resistant towards RNase R treatment. We
229 constructed a minigene expression vector including the two exons of LPAR1 and a portion of the
230 flanking introns (**Fig. 4B**). As with most other AUG circRNA, LPAR1 has no IAE in close proximity, and
231 the mode of biogenesis is therefore currently unclear. To overcome this, we artificially inverted and
232 inserted part of the upstream intron downstream of the splice donor (**Fig. 4B**), which results in clean
233 and efficient expression of circLPAR1 exhibiting RNase R resistance (**Figs. 4C- D**). The protein-coding
234 ability of circLPAR1 was determined by the insertion of an eGFP tag just upstream of the stop codon
235 in the putative ORF (**Fig. 4B**). Here, detection of the tags would only be possible if translation
236 proceeds across the BSJ. First, we tested whether the insertion of eGFP would impede or alter the
237 circularization of circLPAR1. As expected, the circLPAR1-eGFP shows changed migration but remains
238 RNase R resistant (**Fig. 4D**). Then, we over-expressed the untagged and GFP-tagged variants of
239 circLPAR1 in HEK293 cells and performed fluorescent microscopy and western blot analyses,
240 however, in both cases we were unable to obtain any signal (**Figs. 4E- F**). Based on AUG circRNA,
241 circSLC8A1, and the ambiguous AUG circRNA, circCDYL, similar vector-designs were constructed
242 (**Supplementary Figs. 9A-B** and **10A-B**) and effective circRNA production was observed
243 (**Supplementary Figs. 9C-E** and **10C-E**), however, once again no GFP-positive signal by western
244 blotting or fluorescent microscopy was obtained (**Supplementary Figs. 9F-G** and **10F-G**).
245 Collectively, this suggests that these specific circRNAs are not subjected to translation under normal
246 conditions in HEK293 cells.

247 In a parallel experiment using the largeT antigen transformed HEK293T cells, which normally show
248 higher expression of ectopic transgenes, we surprisingly observed a faint GFP-positive band on the
249 western and few GFP positive cells when overexpressing circLPAR1-eGFP. However this was seen

250 both for LPAR1 vectors with or without artificial inverted elements and thus irrespective of
251 circularization (**Supplementary Figs. 11A-E**). Moreover, vector-linearization prior to transfection
252 reduced the GFP output, although this was also observed for the canonical GFP expression vector.
253 Collectively, this indicates that in extreme conditions, exon-repeats are likely produced from 'rolling
254 circle' read-through transcription on plasmid templates (see schematics in **Supplementary Fig. 11F**),
255 and, presumably, this is particularly prevalent in circRNA expression vectors, as the vector-encoded
256 poly(A) signal is situated downstream the SD and therefore subjected to U1-mediated repression
257 (Kaida et al. 2010). Thus, vector-based overexpression may generate false positive protein products
258 from capped mRNA indistinguishable from the predicted circRNA-derived peptides, and conclusions
259 based on ectopic expression setups should be drawn with utmost caution.

260

261 *Putative ORFs are not conserved features in AUG circRNAs*

262 As shown above, AUG circRNAs are generally more conserved across species than other circRNAs
263 (i.e. more often found in mouse as a circRNA). If the functional relevance of these AUG circRNAs is
264 to encode protein, features specific for translation should also exert increased conservational
265 restraint. We focused this analysis on the AUG circRNAs within the top 1000 expressed circRNAs
266 from the ENCODE analysis, and used the AUG-containing exon from 'other circRNA'-associated host
267 genes (termed the 'AUG exon') to include exons with comparable expression level but without any
268 evidence of circularization as a control.

269 In theory, the circular topology allows for infinite ORFs without stop-codons. However, this is only
270 predicted for a very small subset of AUG circRNAs (8% in both human and mouse**, Supplementary**
271 **Figs. 12A-B**). For the remaining circRNAs, the predicted ORF terminates shortly after the BSJ (median
272 length of 10 amino acids (aa) after BSJ, **Supplementary Figs. 12C-D**), which is very close to the
273 expected geometric distribution of stop-codon frequency considering the overall 5'UTR nucleotide
274 composition (**Supplementary Figs. 12E-F**). This suggests that the predicted lengths of the circRNA
275 derived peptides are very close to what would be expected by chance.

276 The mRNA ORFs are typically highly conserved between species. In contrast, the 5'UTRs generally
277 exhibit much lower evolutionary constraints. We compared the overall conservation of 5'UTRs, but

278 only considering the AUG-containing exon (see schematics in **Fig. 5A**, upper panel). Here, AUG
279 circRNAs show a significantly higher cross-species conservation compared to the control AUG exons
280 (**Fig. 5A**). Next, to elucidate whether the increased conservation coincides with a putative ORF from
281 the annotated AUG across the BSJ and into the 5'UTR, we determined the relative conservation of
282 predicted stop-codons. As a positive control, we included the annotated stop-codon from the
283 circRNA-derived host genes (termed 'Host mRNA') in the analysis. Based on phastCons scores
284 obtained from the UCSC genome browser, the relative conservation of stop-codon vs downstream
285 triplet was plotted (**Fig. 5B**). This shows, as expected, a notable conservational enrichment of host
286 mRNA stop-codons, however no significant difference between AUG circRNAs and AUG exons is
287 observed, and not even when focusing the analysis on the conserved subset of AUG circRNAs. In
288 agreement with our analysis of ORF lengths, this suggests that the putative stop-codon sequence is
289 not under evolutionary constraints. Finally, we determined the third-nucleotide (the wobble)
290 conservation relative to the two other nucleotides in every codon within the putative ORF after the
291 BSJ (**Fig. 5C**, schematic). Based on phyloP basewise conservation, the annotated mRNA ORFs show
292 a clear and significant decrease in wobble nucleotide conservation, in accordance with previous
293 analyses (Chamary et al. 2006). However, for the AUG circRNA, again, no differential conservation
294 between the wobble position and the two other bases is observed, not even for the conserved
295 subset of AUG circRNA (**Fig. 5C**), supporting the preliminary conclusion that the coding properties
296 of AUG circRNAs are not conserved. Consistently, similar analyses on the murine repertoire of AUG
297 circRNAs produce almost identical results (**Supplementary Figs. 13A-C**), suggesting little or no
298 preservation of the circRNA-specific ORF, however, the exact peptide sequence encoded after the
299 BSJ could be of less significance and therefore not under evolutionary pressure, but still the protein
300 output could be functionally important.

301

302 *AUG circRNAs are generally not templates for translation*

303 To assess the translational potential of circRNA globally, we took advantage of the wide range of
304 ribosome profiling (RiboSeq) data currently available online: ~500 and ~1300 samples from human
305 and mouse origin (**Supplementary Table 5**). After adapter-trimming, we obtained a total of
306 approximately 22 and 24 billion reads between 25 and 35 nucleotides in length from human and

307 mouse, respectively, consistent with an approximate ribosome footprint of ~30nts (Ingolia et al.
308 2009).

309 In general, the distance from the 5'end of footprinting reads to the ribosome P-site (P-site offset) is
310 12 nts (Ingolia et al. 2009; Bazzini et al. 2014), however, for shorter and longer reads this offset
311 varies (Dunn and Weissman 2016). Moreover, we noted that there was a large degree of p-site offset
312 variation between samples, and consequently, we initially analyzed each sample individually by
313 mapping all the 25-35 nt reads onto the *GAPDH* and *ACTB* (beta-actin) mRNAs. For each read-length,
314 we determined the amount of on-frame P-sites for 12, 13, and 14 nts offsets and the associated p-
315 value by binomial test (see example in **Supplementary Fig. 14**), which was used as a measure of
316 dataset fidelity. Thus, in all samples, the efficiency by which each read length (25-35 nts) is able to
317 demarcate translation from noise was determined (**Supplementary Fig. 15**). Here, for both species,
318 the 28-31 nt reads show the highest abundance and fidelity (**Supplementary Fig. 16**).

319 We then applied the reads to circRNAs using the 5'offset with the lowest p-value in the above
320 analysis. To evaluate translation of circRNA, the BSJ is the only circRNA-specific sequence. Therefore,
321 we concatenated the circRNA exons on all the top 1000 expressed circRNAs from the ENCODE data
322 to display the BSJ in a linear manner compatible with short read mapping. As above, we also included
323 the AUG-containing exon from 'Other circRNA'-derived host genes ('AUG exon'), as well as
324 'ambiguous AUG circRNAs', i.e. circRNAs derived from ambiguous host gene isoforms of which at
325 least one is annotated to contain the AUG start site (see schematics on **Fig. 6A**). By plotting the
326 distribution of reads across the BSJ, only a small fraction of reads spans the BSJ compared to the
327 immediate upstream regions (**Fig. 6B**). However, in contrast to previous reports (Guo et al., 2014;
328 You et al., 2015), there is a notable fraction of BSJ-spanning reads defined here as P-site position
329 from -8 to +6 relative to the BSJ (**Fig. 6B**) comprising a 15 nt stretch (5 codons). To ensure that BSJ-
330 spanning reads are in fact likely derivatives of circRNAs, the reads were re-aligned to an assembled
331 transcriptome allowing one mismatch. Particularly for the human RiboSeq data, this discards the
332 majority of reads spanning the BSJ on the 'AUG Exon' subset. In contrast, almost all reads mapping
333 perfectly to the BSJ of bona fide circRNA have no detectable mRNA alignment (**Supplementary Figs.
334 17A** and **B**). Now, when considering the likelihood of each read to actually derive from translating
335 ribosomes, only ~20% of BSJ-spanning reads are from high quality samples (with fdr<0.01), whereas

336 ~70% of the upstream-derived reads are of high quality (**Fig. 6C**). This suggests that across the BSJ,
337 the quality of ribosome profiling data is of particular high relevance and that noise consumes most
338 of the BSJ-spanning reads. This difference in quality is corroborated in mouse although here the
339 difference in quality is less pronounced (**Supplementary Figs. 18A-C**). Nonetheless, to address
340 translation across the BSJ, we filtered out the low quality reads (fdr>0.01) and used the remaining
341 reads to determine whether phasing in accordance with translation of the putative ORFs is evident.
342 Here, we simply counted the number of reads in-frame and out-of-frame on the 5 codon BSJ-
343 spanning stretch and compared this to a 5 codon stretch immediately upstream the BSJ. For all
344 subtypes of circRNA ('AUG circRNA', 'Ambiguous AUG circRNA', and 'Other circRNA') a roughly equal
345 distribution of reads between all three frames is observed across the BSJ, whereas for the upstream
346 region, approximately 50-60% of reads are in-frame, both for humans and mouse (**Fig. 6D** and
347 **Supplementary Fig. 18D**). This strongly suggests that the AUG circRNAs as a whole, or any of the
348 other circRNAs for that matter, are not subjected to translation as evidenced by RiboSeq analysis.
349 However, it is likely that a small and restricted subset of circRNAs is acting as templates for
350 translation, and therefore the signal from these drown in the noise from others. To evaluate this,
351 we analyzed all top 1000 circRNAs with at least 10 BSJ-spanning RiboSeq reads individually. Here, in
352 humans only circUBXN7 – an ambiguous AUG circRNA for which the annotated host gene has
353 multiple start codons – shows an enrichment of in-frame reads although not significant when
354 evaluating unique reads only (fdr (unique reads)=0.057, fisher's exact test, **Fig. 6E,** unique reads
355 shown in parentheses). Similarly, in mouse, no significant phasing of unique reads is observed across
356 the circRNA BSJ (**Supplementary Fig. 18E**).

357

358 *No circRNA-specific peptides are found in proteomics data*

359 To strengthen the conclusion drawn from the previous analyses, we conducted an extensive mining
360 of proteomics data. Based on the top 1000 expressed circRNAs derived from protein-coding host
361 genes, we predicted the resulting peptide from each circRNA, and compiled an associated database
362 of decoys (**Fig. 7A**). Using the established mass-spec search tool, comet (Eng et al. 2012), we
363 searched through 35 mio. spectral data (see **Supplementary Table 8**) derived from the nci60 cell
364 lines, the human proteome draft map, as well as the human brain proteome. All hits and the

365 corresponding e-values were then assigned to either host gene, circRNA, or decoy. Here, in contrast
366 to host genes, the e-values of circRNA-derived peptides are not significantly different from the
367 decoy peptides (**Fig. 7B** and **Supplementary Fig. 19A**). In fact, using a relaxed fdr cutoff of 0.2, not a
368 single circRNA-derived peptide is identified (**Fig. 7B**), and the AUC in a ROC analysis was for all
369 circRNA subgroups close to 0.5 (**Supplementary Fig. 19B**). This adheres with the above conclusions
370 that circRNAs are unlikely to be substrates for translation. Moreover, we focused specifically on
371 circRNAs shown to produce protein in the literature and subjected these candidates to both RiboSeq
372 and Mass-Spec analysis (summarized in **Supplementary Table 9**). Here, in all cases, no supportive
373 Riboseq reads are found, and not a single circRNA specific peptide with e-values below 0.05 is
374 observed. In fact, for many of the putative peptides derived from circRNAs in Yang et al, 2017, the
375 associated ORFs have no canonical start codon, or the circRNAs are without any reads in the ENCODE
376 RNAseq. This indicates to us that many of these identified peptides are likely false positives.

377

378 *CircRNAs are not enriched with 8mer miRNA target sites*

379 Finally, to determine whether AUG circRNAs show evidence of miRNA *sponging*, we characterized
380 the frequency of miRNA target sites on circRNAs. Consistent with other analyses, we focused on the
381 top 1000 expressed circRNA with splice sites derived from protein-coding genes (i.e. ciRS-7 is not
382 included in this analysis), the AUG exons from 'Other circRNA'-producing host genes, and the unique
383 3'UTRs from the top 1000 expressed circRNA host genes. We conducted an 8mer target site analysis
384 based on all the conserved and confident-assigned miRNAs from miRbase (n=338). Here, we observe
385 for all target categories that the overall numbers of 8mers are enriched compared to expected based
386 on permutations (**Supplementary Fig. 20A**). However, this is also observed for the reverse
387 complement sequence, suggesting that mere nucleotide frequency is not a suitable null hypothesis
388 for in silico target site prediction. Instead, considering the phastCons scores of all predicted target
389 sites, we observe, as expected, that the 3'UTR targets are significantly more conserved compared
390 to the reverse complement and the permutated sequences (**Supplementary Fig. 20B**). This, in fact,
391 is also observed for AUG exon, i.e. non-circularized exons with annotated start codons, and to a
392 lesser degree the AUG circRNAs. This could indicate that preserved miRNA targeting in 5'UTRs is
393 occurring, but not in a circRNA specific manner. To elucidate any putative circRNA sponges, we

394 counted the frequency of all circRNA:miRNA pairs. Here, no circRNAs were observed with more than
395 3 target sites for any particular miRNA, in contrast to ciRS-7 with 59 8mer sites for miR-7. In one
396 outlier instance, however, hsa-miR-342-5p has 23 distinct target sites in the 3'UTR of ANKRD11
397 (uc002fnf.1) (**Supplementary Fig. 20C**). As a last resort, based on the Smith-Waterman alignment
398 score for each putative target site, we sought to disclose whether circRNAs were more prone to
399 miRNA-mediated cleavage than expected by change, however here no significant differences were
400 observed between the actual target sequences and the controls (**Supplementary Fig. 20D**).
401 Conclusively, this aligns with previous studies showing no enrichment of miRNA target sites on
402 circRNAs – except for ciRS-7 (Guo et al. 2014), and suggests that circRNAs are unlikely to serve as
403 miRNA sponges.

404

405 ## Discussion and conclusion

406 Here, by thorough disclosure of the circRNA landscape across human and murine tissues and cell
407 lines, we have shown that a certain subclass of circRNAs, namely the AUG circRNAs, are abundantly
408 expressed and conserved across species. The high abundance of AUG circRNAs is partly reflected in
409 increased circular-to-linear ratio suggesting increased backsplicing efficiency. In fact, for the top
410 1000 expressed circRNAs, there is no significant difference in linear spliced reads between AUG
411 circRNA and Other circRNAs. This indicates that the cross-species conservation is not merely a
412 consequence of abundant host gene expression but may instead reflect sequence preservation of
413 the AUG circRNAs specifically. Moreover, the AUG circRNAs associate with very long flanking introns
414 and are devoid of flanking inverted *Alu* elements in contrast to most other circRNAs. The *Alu*-
415 independent biogenesis is consolidated by analysis of RNAseq on Dhx9 depleted cells and Dhx9 HITS-
416 CLIP, where AUG circRNAs generally are insensitive to Dhx9 perturbation and exhibit reduced Dhx9
417 binding in the flanking regions compared to other circRNAs. This strongly suggests that AUG
418 circRNAs are not aberrant RNA species occasionally escaping the Dhx9-mediated surveillance.
419 CircHIPK3, the most abundant circRNA in humans, has proximal IAE and is considered an *Alu*-
420 dependent circRNA. Our analysis shows that circHIPK3 escape regulation by Dhx9 suggesting *Alu*-
421 independent biogenesis. Moreover, circHipk3 is also highly expressed in mouse, but here devoid of
422 any inverted *Alu*-elements in the flanking regions. Instead, the mouse and human loci share the long

423 flanking introns, suggesting that the biogenesis of circHIPK3 may not rely solely on inverted *Alu*
424 elements in the flanking regions. Here, the combination of long flanking introns and proximal IAE
425 could in part explain the high overall circHIPK3 expression levels in cells and tissue. Thus, we
426 conclude that most AUG circRNAs use *Alu*-independent routes of biogenesis that most likely depend
427 on the length and features of the flanking introns. In *Drosophila*, a well-established circRNA from
428 the laccase locus is flanked by inverted repeats required for biogenesis (Kramer et al. 2015),
429 however consistent with our results, circRNAs from *Drosophila* associate in general with very long
430 introns and without any obvious flanking sequences engaging in intron-intron base-pairing
431 (Westholm et al. 2014). Instead, splicing kinetics and backsplicing seem to be closely related
432 features; It has been shown that increased polymerase speed results in decelerated linear splicing
433 (Carrillo Oesterreich et al. 2016) but stimulates backsplicing (Zhang et al. 2016b). Similarly, depletion
434 of splicing factors increase circRNA production by affecting splicing kinetics negatively (Liang et al.
435 2017). Therefore, we propose that some embodiment of slow splicing kinetics – which may be
436 commonly associated with long introns - is a prerequisite for the production of many circRNAs.
437 While in higher eukaryotes, the duration of splicing does not generally correlate with intron length
438 (Hollander et al. 2016), the duration of intron transcription and the concomitant delay of splicing
439 could in itself be of importance. Whether the intrinsic features of long intron splicing are sufficient
440 for efficient backsplicing or whether certain trans-acting factors are required as well is currently
441 unknown, and as such, the exact production of the most abundant and conserved circRNAs remains
442 elusive.

443 Based on cross-species sequence analysis, the 5'UTR sequences contained within AUG circRNAs are
444 overall more conserved than other non-circular 5'UTR elements. However, based on single
445 nucleotide constraints, this conservation is not due to circRNA-specific ORFs. Consistently, upon
446 overexpression of three different AUG circRNAs, no circRNA-specific peptides were detected.
447 Additionally, thorough and extensive RiboSeq analysis suggests that BSJ-spanning reads are not
448 derived from translating ribosomes assuming that the p-site offsets deduced from house-keeping
449 genes (*GAPDH* and *ACTB*) also apply to circRNA translation. Finally, analysis of 35 mio. peptide
450 spectra from mass spectrometry fails to identify a single circRNA-derived peptide with a fdr below
451 0.2. Collectively, this argues that circRNAs in general and AUG circRNAs specifically are not subjected
452 to translation. It should be emphasized that our analyses solely focus on human and mouse samples

453 and therefore the functional features of circRNA in other organisms, such as *Drosophila*, has not
454 been addressed here. Moreover, our results do not necessarily exclude inefficient translation of
455 circRNAs or restricted translation solely under specific conditions. Consequently, it is possible that
456 published circRNAs, such as circMbl in *Drosophila* (Pamudurti et al. 2017), and circZNF609,
457 circSHPRH, circFBXW7, and circLINC-PINT in human cells (Legnini et al. 2017; Zhang et al. 2018a;
458 Yang et al. 2018; Zhang et al. 2018b), indeed engage in protein production, however, our data
459 strongly point towards translation of circRNAs as being a rare and uncommon process.

460 Since the functional characterization of ciRS-7 as a dedicated miR-7 sponge or regulator, many
461 examples of circRNAs acting as miRNA sponges have been proposed. Apart from the 70+ selectively
462 conserved miR-7 sites on ciRS-7, the miRNA sponge potential seems not to be a conserved feature
463 of circRNAs (Guo et al. 2014). Moreover, stoichiometric analysis of circRNA:miRNA:mRNA ratios
464 suggests little or no overall effect upon circRNA-mediated miRNA inactivation (Denzler et al. 2014;
465 Jens and Rajewsky 2015), highlighting that the notion of circRNAs as miRNA sponges is very
466 controversial. This is consistent with our analysis of putative 8mer target sites. Here, the circRNAs
467 showed no overall enrichment of target site frequency, target site conservation, target site
468 alignment scores, nor any emerging circRNA:miRNA pairs, suggesting that from a bioinformatics
469 point of view, circRNAs do not classify as miRNA regulators.

470 Instead, specifically for the AUG circRNAs, expressing translationally inert canonical start codons in
471 its natural sequence context could be useful in certain scenarios, e.g. as binding platforms for or
472 regulators of translation factors in the cytoplasm, and using the circular topology for this purpose
473 seems plausible, although the same stoichiometric issues as for the miRNA sponge hypothesis may
474 apply here. It is possible that the functional relevance of one particular AUG circRNA is very subtle,
475 but that the accumulated contribution of all the circRNAs are of physiological importance. As such,
476 the highly stable and durable circRNAs could constitute a background of non-responsive RNA
477 entities in the cell to ensure robustness by transiently associating with RNA-binding proteins in the
478 cytoplasm thereby reducing non-specific and potentially detrimental RNA-protein interactions. In
479 any case, future research will undoubtedly shed light on the elusive mechanism by which these
480 highly abundant circRNA species are produced and more interestingly elucidate the functional
481 capabilities of AUG circRNAs.

482

483 ## Materials and Methods

484 *Plasmids*

485 All plasmids were generated by PCR with subsequent restriction digest and ligation into pcDNA3
486 (Invitrogen). Primers are listed in **Supplementary Table 6**.

487 *Cell lines and transfection*

488 HEK293 *Flp-In™ T-Rex™* cells (Invitrogen) or HEK293T cells (ATCC) were used for all experiments.
489 The cells were cultured in Dulbecco's modified Eagle's media (DMEM) with GlutaMAX (Thermo
490 Fischer Scientific) supplemented with 10 % foetal bovine serum (FBS) and 1 %
491 penicillin/streptomycin sulphate. The cells were kept at 37 °C and 5 % CO_2. Transient transfections
492 were carried out using calcium phosphate as transfection reagent using standard procedures or
493 Lipofectamine Reagent 2000 (Invitrogen) accordingly to manufacturers protocol. After 24 hours, the
494 media was changed and 48 hours post transfection cells were harvested either by resuspension in
495 a) 2xSDS loading buffer (for western blotting) or b) TRIzol Reagent (Thermo Fisher Scientific) (for
496 RNA purification) adhering to manufacturer's protocol (see below). For linearization of plasmids
497 prior to transfections, the respective plasmids were digested with FastDigest PvuI (Thermo Fisher
498 Scientific).

499 *Northern blotting*

500 Northern blotting was performed as described in (Hansen 2018a). Briefly, 10 µl RNA (1g/l) and 20
501 µL northern loading buffer [58,8% formamide, 6,5% formaldehyde, ethidium bromide, 1,18% MOPS,
502 bromophenol blue] were mixed and denatured at 65°C for 5 minutes. The RNA was separated by
503 electrophoresis on a 1,2% agarose gel containing 3% formaldehyde and 1x MOPS at 75 V. After
504 electrophoresis, the gel was briefly washed in water and exposed to UV to visualize the EtBr stained
505 rRNA bands. The gel was transferred to a Hybond N+ membrane (GE Healthcare) over-night (O/N)
506 in 10xSSC. Then, the membrane was UV cross-linked and pre-hybridized in Church buffer [0,158 M
507 NaH_2PO_4, 0,342 M Na_2HPO_4, 7 % SDS, 1 mM EDTA, 0.5 % BSA, pH 7.5] for one hour at 55°C and
508 subsequently probed with a 5´ radioactively labelled DNA oligonucleotide (see **Supplementary**

509 **Table 6**) at 55 °C O/N. The next day, the membrane was washed twice in 2 x SSC, 0.1 % SDS for 5
510 minutes and twice in 0,2xSSC, 0,1% SDS for 15 minutes. All washes were carried out at 50°C. Finally,
511 the membrane was exposed on a phosphoimager screen and analyzed using Quantity One ® or
512 Image Lab™ software (Bio Rad). Membranes were stripped in boiling stripping buffer [0,1% SDS, 1
513 mM EDTA].

514 *RNase R*

515 For RNase R experiments, 4 µg RNA was digested with 4 U RNase R (Epicentre) in a total reaction
516 volume of 10 µL for 10 minutes at 37 °C. Then, 20 µL northern loading buffer was added, heated at
517 65°C for 5 minutes before loading on an agarose northern gel, see above.

518 *Western blotting*

519 Cells were harvested in 1xPBS and centrifuged at 1200 rpm at 4°C for 5 min. The supernatant was
520 removed and cell pellet was lysed directly and resuspended in 2xSDS loading buffer [125 mM Tris-
521 HCl, pH 6.8, 20 % glycerol, 5 % SDS, 0,2 M DTT]. After resuspension, the samples were boiled at 95°C
522 for 5 minutes before loading 1 % on a 10 % Tris-Glycine SDS-PAGE gel with 10 µL PageRuler Plus
523 Prestained Protein Ladder (Thermo Scientific). After approximately 1 ½ hours electrophoresis
524 proteins were immobilized on an Immobilon-P Transfer Membrane (EMD Milipore) O/N in a wet-
525 blotting chamber. The next day, the membrane was pre-incubated at RT with 20 % skim milk to
526 block unspecific binding for 1 hr. Then, primary antibody (**Supplementary Table 6**) in blocking
527 solution was added and incubated 1 hour at RT, followed by 1 hour incubation with secondary
528 antibody in blocking solution. After each antibody incubation, the membrane was washed 3x5 min
529 in 1xPBS with 0,05 % Tween 20 and subsequently with 1x5 min wash with 1xPBS. Exposure was
530 performed using SuperSignal West Femto Maximum Sensitivity Substrate kit (Thermo Scientific) and
531 Amersham Hyperfilm ECL (GE Healthcare).

532 *Microscopy*

533 48 hours after transfection, cells were imaged live using phase contrast and fluorescence
534 microscopy with normal FITC filter set (using 1-second exposure, ISO200) on an Olympus IX73
535 microscope. Images were merged using ImageJ.

536 *RNAseq datasets and CircRNA detection*

537 Raw RNA sequencing data was downloaded from the ENCODE Consortium
538 (www.encodeproject.org) or the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) (See
539 **Supplementary Tables 1** and **3**). CircRNA prediction was performed by find_circ, version 1.0
540 (Memczak et al. 2013) and circexplorer2.0.1 (Zhang et al. 2016a) adhering to the recommendation
541 by the authors. For find_circ, an increased stringency threshold was used requiring that both
542 adaptor sequences map with highest possible mapping quality (mapq=40) (Hansen 2018b).
543 Moreover, for find_circ and CIRCexplorer2, only circRNAs supported by at least two reads in a given
544 sample was kept. CircRNAs found by both algorithms with the above-mentioned stringency were
545 used in subsequent analyses. CircRNA expression was based on BSJ-spanning reads according to
546 find_circ quantification. Likewise, the circular-to-linear ratio was determined by the total number
547 of BSJ-spanning reads multiplied by two and divided by the total number of linear spliced reads
548 spanning the upstream and downstream splice sites, respectively, as determined by find_circ. RPM
549 values were calculated for each sample as the number of BSJ-spanning reads divided by the total
550 number of reads. mRNA expression (FPKM values) was quantified using cufflinks (Trapnell et al.
551 2010)**.**

552 *circRNA annotation*

553 Annotation of circRNAs was based on UCSC Genes tracks (hg19 and mm10). First, the annotation
554 database was queried for host genes sharing both the circRNA-specific splice sites. If none were
555 found, genes sharing at least one splice-site were queried, and as a last resort, genes fully covering
556 the circRNA locus were retrieved. Similarly, host gene exons were retrieved from annotated
557 isoforms sharing both circRNA producing splice-sites but omitting first and last exons as well as the
558 exons with splice-sites coinciding with the circRNA, and duplicate exons were discarded. The
559 circRNA subclass, i.e. 'AUG circRNA', 'CDS circRNA' etc, was only determined based on host gene
560 ORF annotation, and if multiple host gene entries were recovered with divergent annotation, the
561 circRNA was categorized as 'Ambiguous'. To guide detection of circRNAs by circExplorer2 and to
562 facilitate proper annotation of known circRNAs, two additional entries were manually added to the
563 gene annotation database (shown in **Supplementary Table 7**). Flanking intron lengths were based
564 on the host gene exon-intron structure immediately upstream and downstream the back-splicing

565 spice sites. In case of multiple isoforms with varying intron length, the mean of all flanking introns
566 found was calculated, whereas circRNAs with no flanking introns annotated were discarded. To
567 extract the distance to nearest flanking inverted *Alu* element, the UCSC Repeatmasker tracks (hg19
568 and mm10) were used. Here, the 20 most proximal but flanking *Alu* elements were retrieved
569 irrespective of intron-exon structure on either side of the circRNA, and based on these, the closest
570 possible inverted pair was determined.

571 *Conservation of circRNA*

572 Human circRNAs were converted to mouse (mm10) coordinates using the UCSC liftover tool. Based
573 on the predicted circRNAs from ENCODE mouse data, and on the UCSC gene annotation database,
574 the coordinates coinciding perfectly with mouse circRNA or annotated splice sites were assessed.
575 Converted coordinates not found as splice sites as well as unmapped coordinated were grouped
576 together as 'non-conserved'.

577 *HITS CLIP analysis*

578 For HITS-CLIP analyses, reads were adaptor-trimmed using trim_galore and barcodes were
579 subsequent removed. The reads were pair-wised mapped on to the human genome (hg19) using
580 bowtie2 using default settings, and mapped reads were extracted using samtools. Reads mapping
581 in the flanking introns within 1000bp of the back-splicing splice-sites of the top 1000 expressed
582 human circRNAs were counted and compared.

583 *Conservation of Open reading frame (ORF)*

584 Conservation analysis was performed for all top 1000 circRNAs with both splice sites annotated in a
585 protein-coding host gene. For 'AUG circRNAs' and 'AUG exons', the ORF was predicted based on the
586 annotated start codon whereas for 'Other circRNAs', the longest finite ORF traversing the BSJ was
587 used. Infinite ORFs were not considered in this analysis. In case of multiple isoforms, i.e. alternative
588 exons within the circRNAs, all isoforms were considered equally in the analysis. The 5'UTR
589 conservation and stop-codon conservation were based on the phastCons scores from 100 species
590 (UCSC) using the hg19 or mm10 reference for human and mouse, respectively. For 5'UTRs, only the
591 AUG-containing exon was analyzed, and for stop-codons, the stop-codon triplet and the immediate
592 downstream triplet irrespective of position was evaluated. For wobble position analysis, only the

593 circRNA-specific ORF within the 5'UTR was analyzed. Here, for each position (1st, 2nd and wobble) in
594 each codon, the PhyloP scores from 100 species (UCSC) was retrieved and analyzed. The ORF lengths
595 were based on number of codons from BSJ to stop, and the geometric distribution of stop codon
596 probability considering the frequency of the individual nucleotides in the 5'UTR region was used to
597 determine the expected ORF lengths.

598 *RiboSeq analysis*

599 Ribosome profiling (RiboSeq) datasets (see **Supplementary Table 5**) were trimmed using
600 trim_galore and only reads between 25 and 35 nucleotides in length were kept and mapped onto
601 *GAPDH* and *ACTB* mRNA (UCSC accessions; uc001qop.2 and uc003sot.4, or uc009dts.2 and
602 uc009ajk.2, for human and mouse, respectively). For each read-length in each RiboSeq sample, an
603 offset of 12, 13 or 14 nucleotides was tested to determine the best possible offsetting based on one-
604 tailed binomial tests, e.g. how many codons exhibit more on-frame than off-frame reads (see
605 **Supplementary Figs. 14-16**). The offset with the lowest mean p-value (obtained from *GAPDH* and
606 *ACTB*) was used for each given read-length in a given sample. The mature sequence of all top 1000
607 circRNAs with both splice sites annotated in a protein-coding host gene were then concatenated to
608 allow mapping across the BSJ. All RiboSeq reads were then bowtie-mapped to the concatenated
609 circRNA sequences allowing no mismatches using the following arguments: *bowtie -S -a -v 0*. The
610 offset from the quality assessment on *GAPDH* and *ACTB* was used to obtain p-site position. P-sites
611 within [-8;6] relative to the BSJ was defined as BSJ-spanning whereas P-sites immediately upstream
612 the SD [-31;-17] was defined as linear upstream reads. Putative BSJ-spanning reads were mapped
613 against an mRNA reference (build on UCSC annotations) with one mismatch tolerance: *bowtie -f -v*
614 *1*, and omitted from downstream analysis if mapped. The annotated AUG was used to predict the
615 circRNA-specific ORF, however in 'Other circRNAs', the longest possible ORF traversing the BSJ was
616 used. Again, infinite ORFs were not considered. P-site positions relative to the predicted frames
617 were counted and analyzed. The statistical significance of the proportion of in-frame reads was
618 determined by Fisher's exact test.

619 *Mass-spec analysis*

620 First, a reference proteome database was compiled. This contained all well-known contaminants,
621 the uniprot protein collection (June 27, 2018, n=73099), and all the putative proteins produced from

622 the circRNAs subjected to RiboSeq analysis. Then, decoys were generated from all the circRNA-
623 derived proteins using the decoyPyRat python script
624 (https://www.sanger.ac.uk/science/tools/decoypyrat) and concatenated to the proteome
625 reference database. Raw MS files (see **Supplementary table 8**) were converted to mgf format using
626 MSconvert (Chambers et al. 2012). Then, the spectra were searched against the reference database
627 using the comet search engine (Eng et al. 2012) adhering to the suggested fixed and variable
628 modifications for each sample. Only the peptides derived from circRNA-based ORFs and the decoys
629 were analysed, and the peptides were stratified by overlap with the Uniprot assembly indicating
630 host gene or circRNA origin. Scores and e-values were extracted and analysed. Fdr values were
631 calculated as the cumulative fraction of decoys ordered by e-value for circRNAs and host genes,
632 respectively.

633

634 circRNA sponge analysis

635 To determine miRNA target sites on circRNAs, all 'confident' human mature miRNA were retrieved
636 from miRBase version 21. Only miRNAs with phastCons scores > 0.95 were kept (n=338). Then 8mer
637 seed-matches (using miRNA positions 2-8 and a downstream A) were searched in the sequence from
638 the top 1000 expressed human circRNAs with both splice-sites annotated in protein-coding host
639 genes. In addition, all unique 3'UTRs from circRNA producing host genes were included. For each
640 analyzed target sequence, the reverse complement sequence as well as 20 random permutations
641 were included in the analysis. The number of putative targets identified where normalized to the
642 expected number of targets based on the nucleotide compositions of target sequence and miRNA
643 seedmatch. Conservation of target sites where based on the phastCons score of the seed-match.
644 Smith-Waterman alignments (using 2, -1, -1 scores for match, mismatch and gap, respectively) were
645 generated between the full miRNA sequence excluding the first nucleotide, and the 8mer target site
646 including an additional 18 nucleotides upstream. Finally, for each target species, all circRNA:miRNA
647 pairs were grouped by counts and summarized.

648

649 *Statistical analyses*

650 All statistical analyses are based on Wilcoxon Rank-sum tests except if explicitly noted otherwise.
651 Fdr-values reflect Benjamini-Hochberg-adjusted p-values, except for the mass-spec analysis (see
652 above).

653

654 ## Figure Legends

655 **Figure 1: Abundant circRNAs. A**) Venndiagram showing the number of exotic (orange) and shared
656 (blue) circRNAs found by find_circ and circexplorer2 algorithms in the ENCODE datasets. **B**)
657 Smoothed fraction of shared circRNAs found by find_circ and circexplorer2 as a function of ranked
658 expression. **C**) Scatterplot depicting the number of backsplice junction (BSJ)-spanning reads
659 obtained from find_circ and circexplorer2 across all the samples analyzed. The points are color-
660 coded as shared (blue) or exotic (orange). **D**) The alpha frequency of the top 10 most commonly
661 found alpha circRNAs as well as the frequency of being an abundant circRNA (i.e. one of the top 10
662 expressed circRNAs in a sample) are plotted as a stacked barplot. **E**) Schematic illustration of the
663 flanking intron length and inverted *Alu* element (IAE). **F**) Boxplot comparing the distance to inverted
664 repeat element and flanking intron length for circRNAs in general (n=81589), host gene exons
665 (n=131002) and the top 10 alpha circRNAs. *, $p < 0.05$; **, $p < 0.01$, Wilcoxon rank-sum test. **G**)
666 Density-colored scatterplot showing relationship between IAE and flanking intron length for all
667 circRNAs (left) and host gene exons (right). The top 10 alpha circRNAs are highlighted to the left.

668 **Figure 2: The AUG circRNAs. A**) Schematics showing circRNA annotation. The genic features are not
669 drawn to scale. **B**) Frequency of circRNA- annotations for either all circRNA, the top 1000, top 100,
670 top 10 circRNAs based on overall expression (RPM), and the top 10 alpha circRNAs, color-coded as
671 denoted. P-values are calculated using Fisher's exact test. **C**) Smoothed relationship between
672 circRNA expression and frequency of AUG containing circRNA and conservation to mouse, i.e. found
673 as circRNA in mouse ENCODE RNAseq. **D**) Based on the top 1000 expressed circRNAs, the fraction
674 of circRNA coordinates found as circRNA in the mouse ENCODE data (Conserved circRNA) stratified
675 by annotation is shown. Moreover, the fraction of conserved splice sites, i.e corresponding to
676 annotated splice sites in mouse, conserved splice donor (SD), conserved splice acceptor (SA), and
677 sites without annotation are depicted. P-value is calculated using Fisher's exact test using conserved

678 circRNAs from AUG and non-AUG stratification. **E-F**) Inverted *Alu* element (IAE) distance (E) and
679 flanking intron length (F) for circRNAs stratified by annotation or by conservation, as well as host
680 gene exons. For flanking intron length (F), only circRNAs and exons with annotated up- and
681 downstream introns were included in the analysis. P-values are based on Wilcoxon rank-sum tests,
682 and where 'AUG circRNAs' are compared to 'Other circRNAs' and 'Ambiguous circRNAs', only the
683 highest obtained p-value is denoted.

684 **Figure 3: AUG circRNAs are DHX9-resistant. A**) Volcanoplot on top 1000 expressed circRNAs from
685 Aktas *et al*, 2017, showing circRNA deregulation upon DHX9 knockdown color-coded by fold-change
686 significance (fdr < 0.05). The top 10 expressed circRNAs are highlighted and color-coded by
687 annotation. **B**) Scatterplot showing flanking intron length by flanking *Alu* distance. Here, as in (A)
688 the top 10 expressed circRNAs are highlighted. **C**) Boxplot showing the distribution of IAE distance
689 and flanking intron length for circRNAs stratified by Dhx9 sensitivity. **D**) Frequency of circRNA-
690 annotations for circRNAs resistant or sensitive towards Dhx9 knockdown. P-value is calculated by
691 Fisher's exact test using AUG and non-AUG stratification. **E**) Within the top 1000 expressed circRNA,
692 the fraction of Dhx9-sensitive species are grouped by annotation and plotted. **F**) Cumulative plot
693 showing the number of Dhx9 HITS-CLIP reads in the flanking vicinity (within 1kb upstream and
694 downstream of the SA and SD, respectively) of circRNAs. Here, the circRNAs were stratified by either
695 genic annotation or Dhx9 sensitivity. P-values are calculated by Wilcoxon rank-sum test, and refer
696 to the 'AUG circRNA' vs 'Other circRNA' and 'DHX9 sensitive' vs 'DHX9 resistant' subgroups.

697 **Figure 4: No evidence of circLPAR1 translation. A**) Genomic representation of the *LPAR1* host gene
698 locus. The exons are not drawn to scale. **B**) Schematic representation of expression vectors
699 comprising the CMV promoter, exons 2 and 3 known to circularize, the putative circRNA-specific
700 stop-codon (cSTOP), the insertion of eGFP ORF, the flanking regions (divergent arrows indicate
701 artificially introduced inverted element) and the BGH pA signal. **C-D**) Northern blot analysis of total
702 RNA from ectopic overexpression of circLPAR1 vectors as denoted in HEK293 cells (C) or RNA with
703 or without RNase R treatment (D). The membranes were probed for circLPAR1 as denoted to the
704 right (top panels), and 18S (obtained by ethidium bromide stain) serves as loading and RNase R
705 control (bottom panels). **E**) Western blot showing GFP expression in HEK293 cells transfected with

706 positive control (pcDNA-eGFP) or circLPAR1-eGFP fusion. **F**) Merged phase contrast and GFP
707 fluorescence images (PC/GFP) obtained from HEK293 cells transfected with vectors as denoted.

708 **Figure 5: No evolutionary preservation of Open-Reading-Frames in AUG circRNAs. A**) PhastCons
709 analysis of 5'UTRs within the AUG-containing exon performed on AUG-containing exons in
710 conserved and non-conserved AUG circRNAs, as well as AUG-containing exons from non-circular
711 AUG exons. The 5'UTRs (transparent red and transparent grey) and representative phastCons tracks
712 are depicted for 'AUG circRNAs' and 'AUG exons', respectively, in the above schematics. **B**)
713 PhastCons scores of putative circRNA-derived stop-codon positions were extracted and compared
714 to the PhastCons scores of the immediate downstream triplet (+3) to obtain a measure of selective
715 conservation (Relative Stop-codon conservation). The analysis was stratified by annotation as in (A)
716 and visualized as a cumulative fraction plot. The stop-codon and +3 triplet is exemplified by
717 circSLC8A1 to the right. Similar analysis on bona-fide stop-codons within host gene ORFs is included
718 (black line). **C**) PhyloP analysis of single-position conservation for 1st, 2nd and wobble-position for
719 bona-fide ORFs within host genes and putative ORFs after BSJ as exemplified to the right by
720 circSLC8A1. N denotes number of codons analyzed.

721 **Figure 6: Ribosome profiling reads across BSJ. A**) Schematics showing circRNA annotation of 'AUG
722 circRNA', 'Ambiguous circRNA', and 'Other circRNA'. **B**) Based on Ribosome profiling datasets, the
723 number of ribosome P-sites around the backsplice junction (BSJ) were counted for each subclass of
724 circRNA ('AUG circRNA', 'Ambiguous AUG circRNAs', and 'Other circRNA', see text for more detail).
725 The AUG-containing exon from non-AUG ('AUG exon') circRNA host genes. The plot is color-scaled
726 according to the associated read-class p-value (See **Supplementary Figs. 14** and **15**). The grey box
727 denotes the defined P-site position of BSJ-spanning reads, from pos -8 to +6 relative to the BSJ. **C**)
728 Based on all BSJ-spanning reads (P-sites from -8 to +6 relative to BSJ) and upstream reads (-31 to –
729 17 relative to BSJ), the p-value distribution based on RiboSeq quality assessment is shown. **D**)
730 Phasing of reads across BSJ. Here, based solely on high quality reads ($p<0.01$), the fraction of P-sites
731 in-frame and out-of-frame across the BSJ (-8 to +6) are shown for each subclass of circRNA. The
732 number below the plot represent total number of reads analyzed, whereas the numbers inside the
733 plot reflect total or unique (in parenthesis) counts within each frame. **E**) As in (D), but for each

individual circRNA with 10+ reads across the BSJ. P-values in (D) and (E) represent one-tailed exact binomial tests.

Figure 7: circRNA-derived peptides. A) Schematics outlining the proteomics analysis pipeline. For the top 1000 expressed circRNAs, the putative ORFs were deduced. From this, a reference proteome was established comprising the entire circRNA-derived peptides as well as a matched list of decoys. Then, proteomics data was queried against the reference database using comet and hits were stratified by 'circRNA', 'Host gene' and 'Decoy'. **B)** Boxplot showing the distribution of obtained e-values from 'circRNA', 'Host gene' or 'Decoy' subclassified into circRNA annotation as shown. **C)** Cumulative sum of peptide hits with incrementing false discovery rate (FDR) cutoffs. FDRs are calculated as the cumulative fraction of decoys on e-value-sorted data with circRNAs and Host genes analyzed separately.

Acknowledgements

We would like to thank Karoline Krogh Ebbesen for the insightful discussions and critical reading of the manuscript. This work was supported by the Novo Nordisk Foundation (NNF16OC0019874 to T.B.H)

References

Aktaş T, Avşar Ilık İ, Maticzka D, Bhardwaj V, Pessoa Rodrigues C, Mittler G, Manke T, Backofen R, Akhtar A. 2017. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544**: 115–119. http://www.nature.com/doifinder/10.1038/nature21715 (Accessed September 5, 2017).

Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. 2014. circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell* **56**: 55–66.

760 Bachmayr-Heyda A, Reiner AT, Auer K, Sukhbaatar N, Aust S, Bachleitner-Hofmann T, Mesteri I,
761 Grunt TW, Zeillinger R, Pils D. 2015. Correlation of circular RNA abundance with proliferation-
762 -exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human
763 tissues. *Sci Rep* **5**: 8057. http://www.ncbi.nlm.nih.gov/pubmed/25624062.

764 Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**: 215–233.

765 Bazzini AA, Johnstone TG, Christiano R, MacKowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee
766 MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using
767 ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981–993.

768 Böhmdorfer G, Wierzbicki AT. 2015. Control of Chromatin Structure by Long Noncoding RNA.
769 *Trends Cell Biol* **25**: 623–32. http://www.ncbi.nlm.nih.gov/pubmed/26410408 (Accessed
770 September 1, 2017).

771 Carrillo Oesterreich F, Herzel L, Straube K, Hujer K, Howard J, Neugebauer KM. 2016. Splicing of
772 Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell*.

773 Chaiteerakij R, Zhang X, Addissie BD, Essa A, Harmsen WS, Theobald PJ, Peters BE, Joseph G, Ward
774 MM, Giama NH, et al. 2017. Circular RNA MTO1 acts as the sponge of miR-9 to suppress
775 hepatocellular carcinoma progression. *Hepatology* 2–43.

776 Chamary J V., Parmley JL, Hurst LD. 2006. Hearing silence: Non-neutral evolution at synonymous
777 sites in mammals. *Nat Rev Genet* **7**: 98–108.

778 Chambers MC, MacLean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt
779 B, Egertson J, et al. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nat*
780 *Biotechnol*.

781 Cocquerelle C, Mascrez B, Hetuin D, Bailleul B, Hétuin D, Bailleul B. 1993. Mis-splicing yields
782 circular RNA molecules. *FASEB J* **7**: 155–160.

783 Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, Roslan S, Schreiber AW,
784 Gregory PA, Goodall GJ. 2015. The RNA binding protein quaking regulates formation of
785 circRNAs. *Cell* **160**: 1125–1134.

786 Denzler R, Agarwal V, Stefano J, Bartel DP, Stoffel M. 2014. Assessing the ceRNA Hypothesis with
787 Quantitative Measurements of miRNA and Target Abundance. *Mol Cell* **54**: 766–776.

788 Dunn JG, Weissman JS. 2016. Plastid: Nucleotide-resolution analysis of next-generation
789 sequencing and genomics data. *BMC Genomics* **17**.

790 Ebbesen KK, Kjems J, Hansen TB. 2016. Circular RNAs: Identification, biogenesis and function.
791 *Biochim Biophys Acta* **1859**: 163–168.

792 ENCODE Project Consortium TEP. 2012. An integrated encyclopedia of DNA elements in the human
793 genome. *Nature* **489**: 57–74. http://www.ncbi.nlm.nih.gov/pubmed/22955616 (Accessed
794 September 1, 2017).

795 Eng JK, Jahan TA, Hoopmann MR. 2012. Comet: an open source tandem mass spectrometry
796 sequence database search tool. *Proteomics*.

797 Guo JU, Agarwal V, Guo H, Bartel DP. 2014. Expanded identification and characterization of
798 mammalian circular RNAs. *Genome Biol* **15**: 409–014–0409–z.

799 Hansen TB. 2018a. Characterization of circular RNA concatemers. In *Methods in Molecular Biology*,
800 Vol. 1724 of, pp. 143–157 http://link.springer.com/10.1007/978-1-4939-7562-4_12 (Accessed
801 January 26, 2018).

802 Hansen TB. 2018b. Improved circRNA Identification by Combining Prediction Algorithms. *Front Cell*
803 *Dev Biol* **6**. http://journal.frontiersin.org/article/10.3389/fcell.2018.00020/full.

804 Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. 2013a. Natural
805 RNA circles function as efficient microRNA sponges. *Nature* **495**: 384–388.
806 http://www.ncbi.nlm.nih.gov/pubmed/23446346.

807 Hansen TB, Kjems J, Damgaard CK. 2013b. Circular RNA and miR-7 in Cancer. *Cancer Res* **73**: 5609–
808 5612.

809 Hansen TB, Veno MT, Damgaard CK, Kjems J, Venø MT, Damgaard CK, Kjems J. 2015. Comparison
810 of circular RNA prediction tools. *Nucleic Acids Res* **gkv1458**.

811 Hansen TB, Wiklund ED, Bramsen JB, Villadsen SB, Statham AL, Clark SJ, Kjems J. 2011. miRNA-
812 dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA.
813 *EMBO J* **30**: 4414–4422. http://www.ncbi.nlm.nih.gov/pubmed/21964070.

814 Hollander D, Naftelberg S, Lev-Maor G, Kornblihtt AR, Ast G. 2016. How Are Short Exons Flanked
815 by Long Introns Defined and Committed to Splicing? *Trends Genet* **32**: 596–606.
816 http://linkinghub.elsevier.com/retrieve/pii/S0168952516300798 (Accessed September 5,
817 2017).

818 Huang C, Liang D, Tatomer DC, Wilusz JE. 2018. A length-dependent evolutionarily conserved
819 pathway controls nuclear export of circular RNAs. *Genes Dev*.
820 http://www.ncbi.nlm.nih.gov/pubmed/29773557 (Accessed May 23, 2018).

821 Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of
822 translation with nucleotide resolution using ribosome profiling. *Science (80-)* **324**: 218–223.

823 Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. 2013.
824 Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**: 141–57.
825 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3543092&tool=pmcentrez&rend
826 ertype=abstract.

827 Jens M, Rajewsky N. 2015. Competition between target sites of regulators shapes post-
828 transcriptional gene regulation. *Nat Rev Genet* **16**: 113–126.

829 Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-
830 mRNAs from premature cleavage and polyadenylation. *Nature* **468**: 664–668.
831 http://www.ncbi.nlm.nih.gov/pubmed/20881964 (Accessed May 23, 2018).

832 Kramer MC, Liang D, Tatomer DC, Gold B, March ZM, Cherry S, Wilusz JE. 2015. Combinatorial
833 control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and SR proteins.
834 *Genes Dev* **29**: 2168–2182.

835 Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, Fatica A, Santini T,
836 Andronache A, Wade M, et al. 2017. Circ-ZNF609 Is a Circular RNA that Can Be Translated and
837 Functions in Myogenesis. *Mol Cell* **66**: 22–37.e9.

838 Li X, Liu C-X, Xue W, Zhang Y, Jiang S, Yin Q-F, Wei J, Yao R-W, Yang L, Chen L-L. 2017. Coordinated
839 circRNA Biogenesis and Function with NF90/NF110 in Viral Infection. *Mol Cell* **67**: 214–227.e7.
840 http://www.ncbi.nlm.nih.gov/pubmed/28625552 (Accessed September 1, 2017).

841 Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, et al. 2015. Exon-intron
842 circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* **22**: 256–264.
843 http://www.nature.com/nsmb/journal/v22/n3/full/nsmb.2959.html%5Cnhttp://www.nature.
844 com/nsmb/journal/v22/n3/pdf/nsmb.2959.pdf.

845 Liang D, Tatomer DC, Luo Z, Wu H, Yang L, Chen L-L, Cherry S, Wilusz JE. 2017. The Output of
846 Protein-Coding Genes Shifts to Circular RNAs When the Pre-mRNA Processing Machinery Is
847 Limiting. *Mol Cell* **68**: 940–954.e3. http://www.ncbi.nlm.nih.gov/pubmed/29174924
848 (Accessed May 23, 2018).

849 Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen
850 LH, Munschauer M, et al. 2013. Circular RNAs are a large class of animal RNAs with regulatory
851 potency. *Nature* **495**: 333–338. http://www.ncbi.nlm.nih.gov/pubmed/23446348.

852 Noh JH, Kim KM, McClusky WG, Abdelmohsen K, Gorospe M. 2018. Cytoplasmic functions of long
853 noncoding RNAs. *Wiley Interdiscip Rev RNA*.

854 Pamudurti NR, Bartok O, Jens M, Ashwal-Fluss R, Stottmeister C, Ruhe L, Hanan M, Wyler E, Perez-
855 Hernandez D, Ramberger E, et al. 2017. Translation of CircRNAs. *Mol Cell* **66**: 9–21.e7.
856 http://www.ncbi.nlm.nih.gov/pubmed/28344080 (Accessed August 31, 2017).

857 Peng L, Chen G, Zhu Z, Shen Z, Du C, Zang R, Su Y, Xie H, Li H, Xu X, et al. 2016. Circular RNA ZNF609
858 functions as a competitive endogenous RNA to regulate AKT3 expression by sponging miR-
859 150-5p in Hirschsprung's disease. *Oncotarget* **5**: 808–818.
860 http://www.oncotarget.com/abstract/13656.

861 Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Hanan M, Behm M, Bartok O, Ashwal-Fluss
862 R, Herzog M, et al. 2015. Circular RNAs in the Mammalian Brain Are Highly Abundant,
863 Conserved, and Dynamically Expressed. *Mol Cell* **58**: 870–885.

864 Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. 2012. Circular RNAs are the predominant

865 transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* **7**: e30733.

866 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ,
867 Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated
868 transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

869 Veno MT, Hansen TB, Veno ST, Clausen BH, Grebing M, Finsen B, Holm IE, Kjems J, Venø MT,
870 Hansen TB, et al. 2015. Spatio-temporal regulation of circular RNA expression during porcine
871 embryonic brain development. *Genome Biol* **16**: 243–245.
872 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4635978&tool=pmcentrez&rend
873 ertype=abstract.

874 Wang Y, Wang Z. 2015. Efficient backsplicing produces translatable circular mRNAs. *RNA* **21**: 172–
875 179.

876 Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, Lai EC.
877 2014. Genome-wide analysis of drosophila circular RNAs reveals their structural and
878 sequence properties and age-dependent neural accumulation. *Cell Rep* **9**: 1966–1980.

879 Yang Y, Fan X, Mao M, Song X, Wu P, Zhang Y, Jin Y, Yang Y, Chen L-L, Wang Y, et al. 2017.
880 Extensive translation of circular RNAs driven by N6-methyladenosine. *Cell Res* **27**: 626–641.
881 http://www.ncbi.nlm.nih.gov/pubmed/28281539 (Accessed August 31, 2017).

882 Yang Y, Gao X, Zhang M, Yan S, Sun C, Xiao F, Huang N, Yang X, Zhao K, Zhou H, et al. 2018. Novel
883 Role of FBXW7 Circular RNA in Repressing Glioma Tumorigenesis. *J Natl Cancer Inst*.

884 Zaphiropoulos PG. 1997. Exon skipping and circular RNA formation in transcripts of the human
885 cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in
886 testis. *Mol Cell Biol* **17**: 2985–2993.

887 Zhang M, Huang N, Yang X, Luo J, Yan S, Xiao F, Chen W, Gao X, Zhao K, Zhou H, et al. 2018a. A
888 novel protein encoded by the circular form of the SHPRH gene suppresses glioma
889 tumorigenesis. *Oncogene*.

890 Zhang M, Zhao K, Xu X, Yang Y, Yan S, Wei P, Liu H, Xu J, Xiao F, Zhou H, et al. 2018b. A peptide

891 encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in
892 glioblastoma. *Nat Commun* **9**: 4475. http://www.nature.com/articles/s41467-018-06862-2
893 (Accessed November 23, 2018).

894 Zhang X-O, Dong R, Zhang Y, Zhang J-L, Luo Z, Zhang J, Chen L-L, Yang L. 2016a. Diverse alternative
895 back-splicing and alternative splicing landscape of circular RNAs. *Genome Res* **26**: 1277–87.
896 http://genome.cshlp.org/lookup/doi/10.1101/gr.202895.115 (Accessed September 1, 2017).

897 Zhang XO, Wang H Bin, Zhang Y, Lu X, Chen LL, Yang L. 2014. Complementary sequence-mediated
898 exon circularization. *Cell* **159**: 134–147.

899 Zhang Y, Xue W, Li X, Zhang J, Chen S, Zhang JL, Yang L, Chen LL. 2016b. The Biogenesis of Nascent
900 Circular RNAs. *Cell Rep*.

901 Zheng Q, Bao C, Guo W, Li S, Chen J, Chen B, Luo Y, Lyu D, Li Y, Shi G, et al. 2016. Circular RNA
902 profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple
903 miRNAs. *Nat Commun* **7**: 11215.
904 http://www.nature.com/doifinder/10.1038/ncomms11215%5Cnhttp://www.nature.com/nco
905 mms/2016/160406/ncomms11215/full/ncomms11215.html.

906

Figure 1

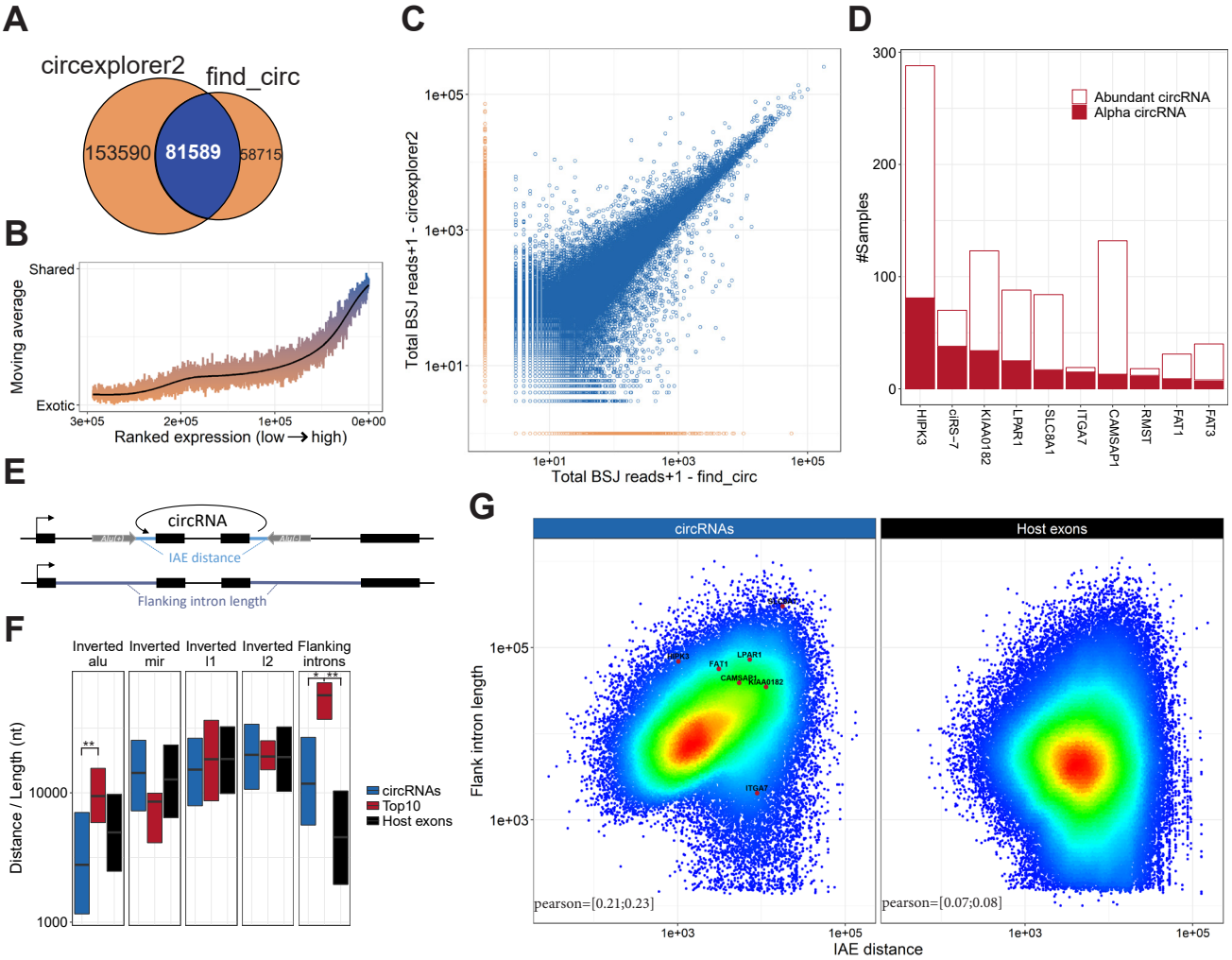

A
circexplorer2
find_circ
153590
81589
58715
B
Shared
Exotic
Moving average
3e+05
2e+05
1e+05
0e+00
Ranked expression (low → high)
C
Total BSJ reads+1 - circexplorer2
Total BSJ reads+1 - find_circ
1e+01
1e+03
1e+05
D
Abundant circRNA
Alpha circRNA
#Samples
300
200
100
0
HIPK3
ciRS-7
KIAA0182
LPAR1
SLC8A1
ITGA7
CAMSAP1
RMST
FAT1
FAT3
E
circRNA
IAE distance
Flanking intron length
F
Inverted alu
Inverted mir
Inverted l1
Inverted l2
Flanking introns
Distance / Length (nt)
10000
1000
circRNAs
Top10
Host exons
G
circRNAs
Host exons
Flank intron length
IAE distance
pearson=[0.21;0.23]
pearson=[0.07;0.08]

Figure 2

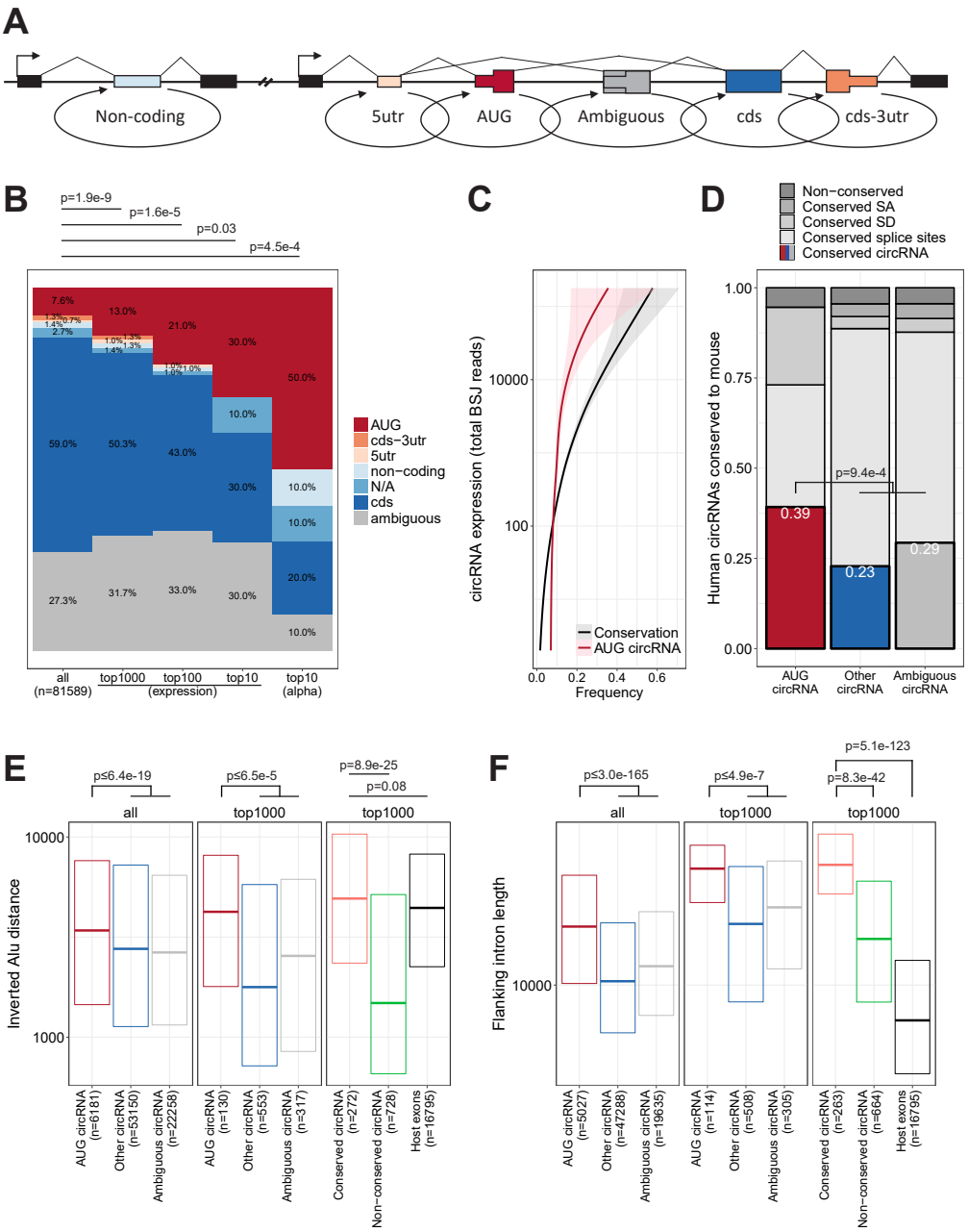

A
Non-coding
5utr
AUG
Ambiguous
cds
cds-3utr
B
p=1.9e-9
p=1.6e-5
p=0.03
p=4.5e-4
AUG
cds−3utr
5utr
non−coding
N/A
cds
ambiguous
all (n=81589)
top1000
top100
top10
(expression)
top10 (alpha)
C
circRNA expression (total BSJ reads)
Frequency
Conservation
AUG circRNA
D
Non−conserved
Conserved SA
Conserved SD
Conserved splice sites
Conserved circRNA
Human circRNAs conserved to mouse
p=9.4e-4
0.39
0.23
0.29
AUG circRNA
Other circRNA
Ambiguous circRNA
E
p≤6.4e-19
p≤6.5e-5
p=8.9e-25
p=0.08
all
top1000
top1000
Inverted Alu distance
AUG circRNA (n=6181)
Other circRNA (n=53150)
Ambiguous circRNA (n=22258)
AUG circRNA (n=130)
Other circRNA (n=553)
Ambiguous circRNA (n=317)
Conserved circRNA (n=272)
Non−conserved circRNA (n=728)
Host exons (n=16795)
F
p≤3.0e-165
p≤4.9e-7
p=5.1e-123
p=8.3e-42
all
top1000
top1000
Flanking intron length
AUG circRNA (n=5027)
Other circRNA (n=47288)
Ambiguous circRNA (n=19635)
AUG circRNA (n=114)
Other circRNA (n=508)
Ambiguous circRNA (n=305)
Conserved circRNA (n=263)
Non−conserved circRNA (n=664)
Host exons (n=16795)

Figure 3

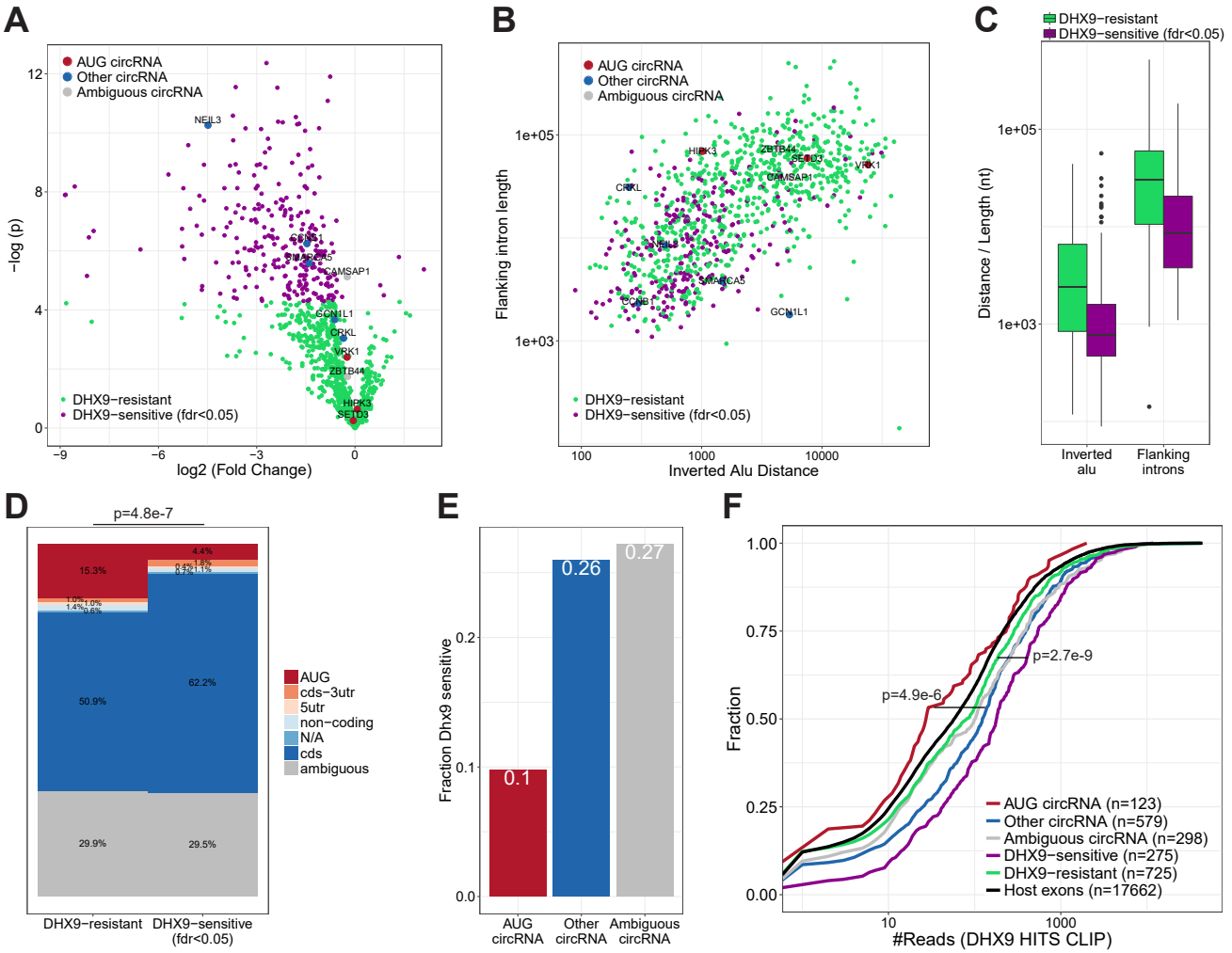

A
AUG circRNA
Other circRNA
Ambiguous circRNA
NEIL3
CAMSAP1
GCN1L1
CRKL
VRK1
ZBTB44
HIPK3
SETD3
DHX9−resistant
DHX9−sensitive (fdr<0.05)
−log (p)
log2 (Fold Change)
B
AUG circRNA
Other circRNA
Ambiguous circRNA
HIPK3
ZBTB44
SETD3
VRK1
CAMSAR1
CRKL
NEIL3
SMARCA5
CCNB1
GCN1L1
DHX9−resistant
DHX9−sensitive (fdr<0.05)
Flanking intron length
Inverted Alu Distance
C
DHX9−resistant
DHX9−sensitive (fdr<0.05)
Distance / Length (nt)
Inverted alu
Flanking introns
D
p=4.8e-7
15.3%
4.4%
50.9%
62.2%
29.9%
29.5%
AUG
cds−3utr
5utr
non−coding
N/A
cds
ambiguous
DHX9−resistant
DHX9−sensitive (fdr<0.05)
E
0.1
0.26
0.27
Fraction Dhx9 sensitive
AUG circRNA
Other circRNA
Ambiguous circRNA
F
p=4.9e-6
p=2.7e-9
AUG circRNA (n=123)
Other circRNA (n=579)
Ambiguous circRNA (n=298)
DHX9−sensitive (n=275)
DHX9−resistant (n=725)
Host exons (n=17662)
Fraction
#Reads (DHX9 HITS CLIP)

Figure 4

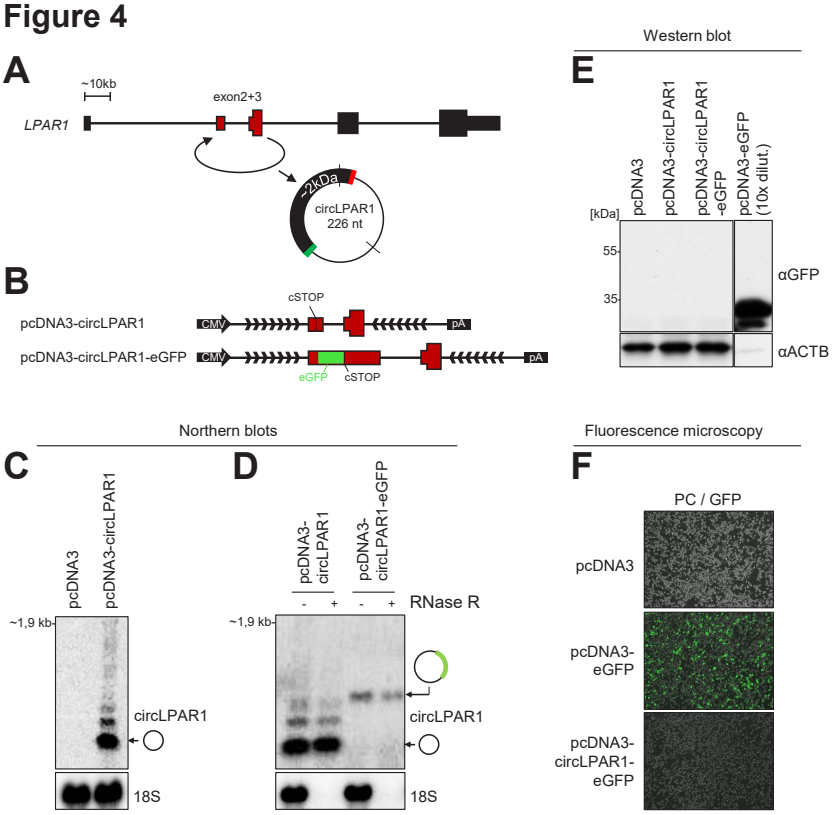

A
~10kb
exon2+3
LPAR1
~2kDa
circLPAR1
226 nt
B
cSTOP
pcDNA3-circLPAR1
CMV
pA
pcDNA3-circLPAR1-eGFP
eGFP
cSTOP
Western blot
E
pcDNA3
pcDNA3-circLPAR1
pcDNA3-circLPAR1 -eGFP
pcDNA3-eGFP (10x dilut.)
[kDa]
55
35
αGFP
αACTB
Northern blots
C
pcDNA3
pcDNA3-circLPAR1
~1,9 kb
circLPAR1
18S
D
pcDNA3-circLPAR1
pcDNA3-circLPAR1-eGFP
RNase R
-
+
-
+
~1,9 kb
circLPAR1
18S
Fluorescence microscopy
F
PC / GFP
pcDNA3
pcDNA3-eGFP
pcDNA3-circLPAR1-eGFP

Figure 5

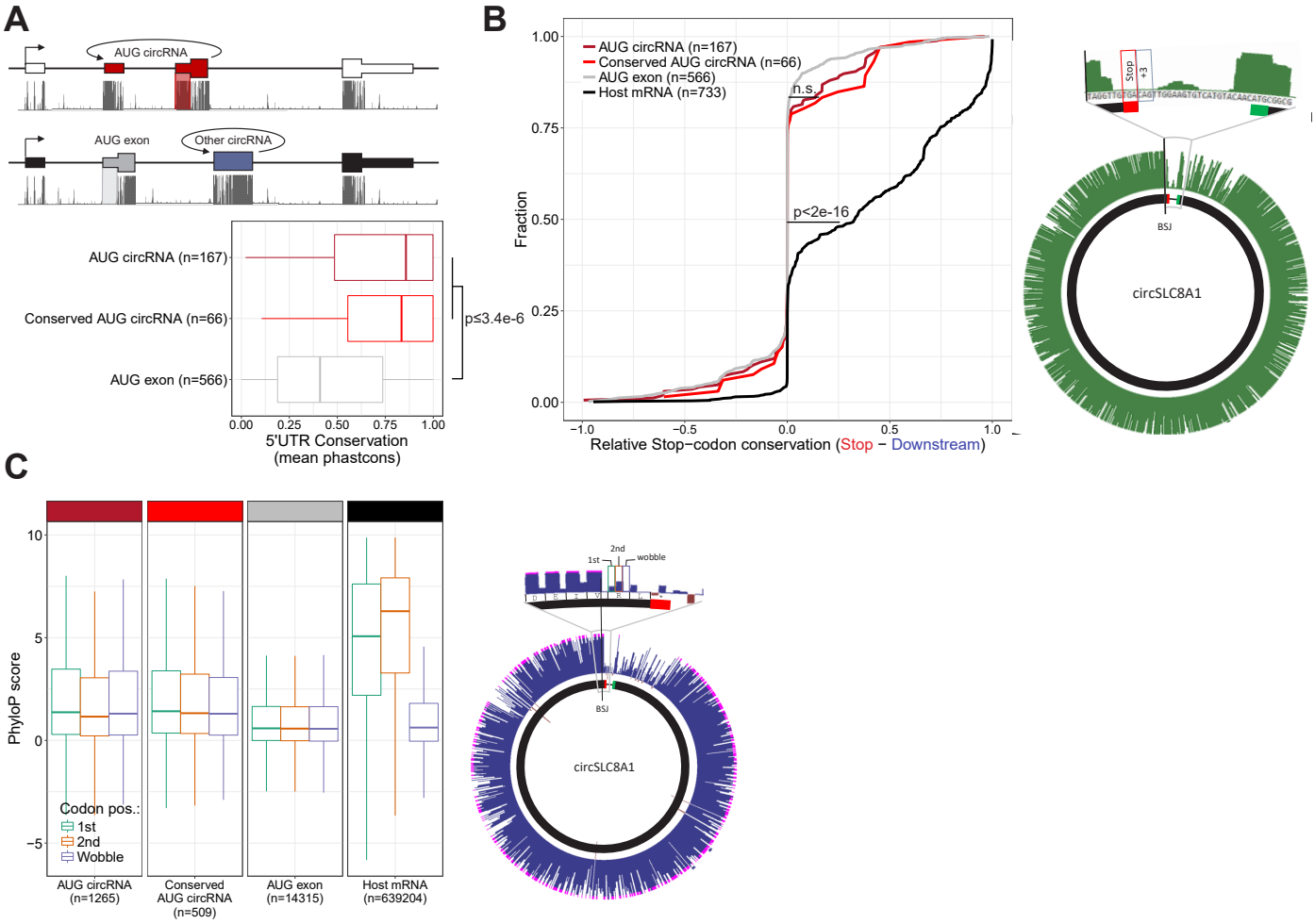

A
AUG circRNA
AUG exon
Other circRNA
AUG circRNA (n=167)
Conserved AUG circRNA (n=66)
AUG exon (n=566)
p≤3.4e-6
0.00
0.25
0.50
0.75
1.00
5'UTR Conservation
(mean phastcons)
B
AUG circRNA (n=167)
Conserved AUG circRNA (n=66)
AUG exon (n=566)
Host mRNA (n=733)
n.s.
p<2e-16
Fraction
Relative Stop−codon conservation (Stop − Downstream)
Stop
+3
BSJ
circSLC8A1
C
PhyloP score
Codon pos.:
1st
2nd
Wobble
AUG circRNA
(n=1265)
Conserved
AUG circRNA
(n=509)
AUG exon
(n=14315)
Host mRNA
(n=639204)
2nd
1st
wobble
BSJ
circSLC8A1

Figure 6

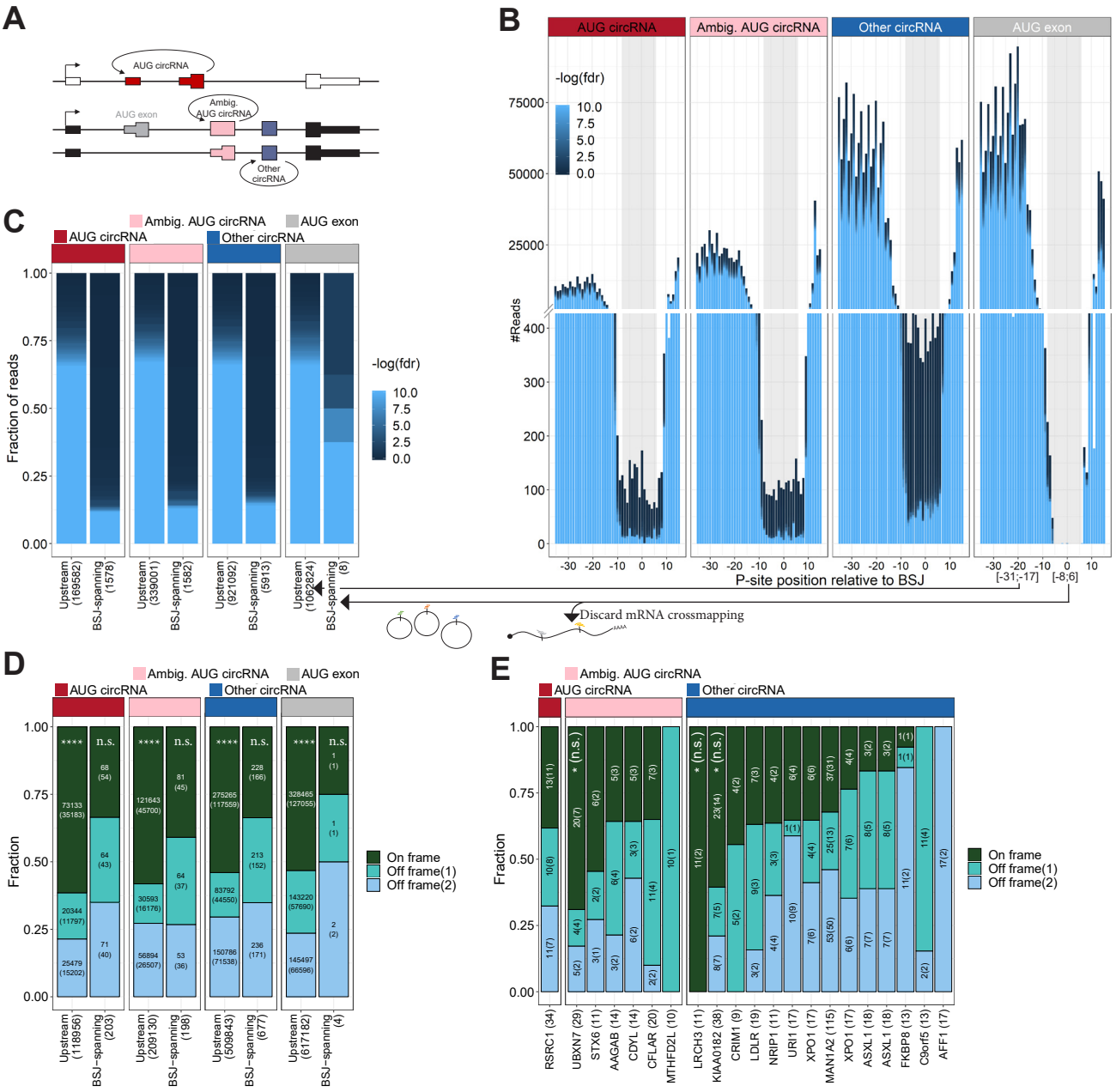
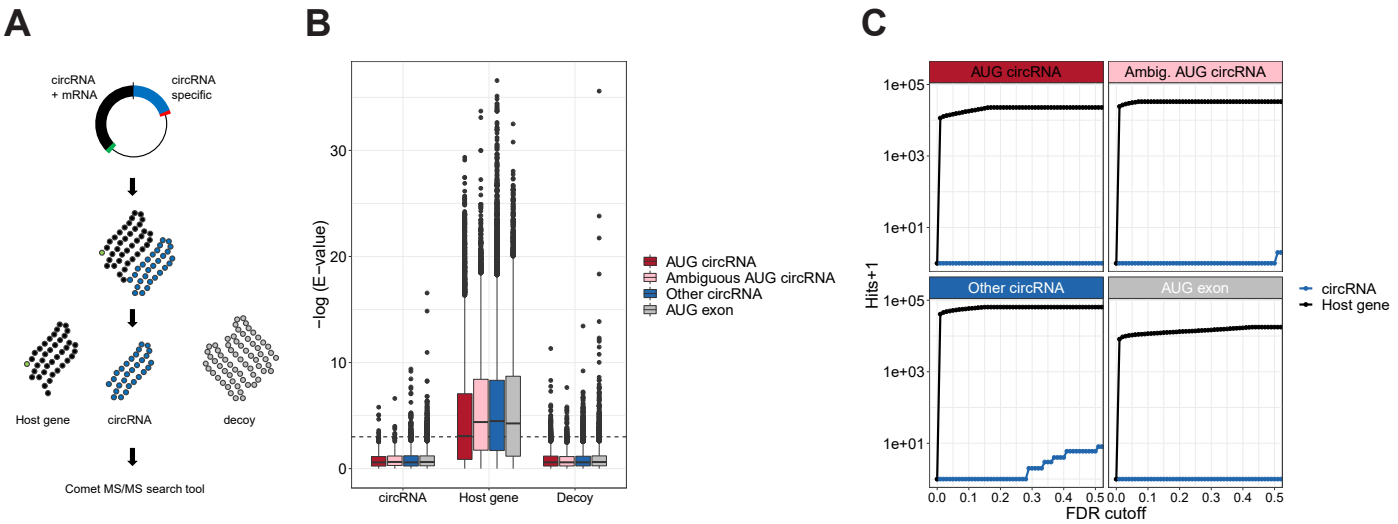

A
AUG circRNA
AUG exon
Ambig. AUG circRNA
Other circRNA
B
AUG circRNA
Ambig. AUG circRNA
Other circRNA
AUG exon
-log(fdr)
10.0
7.5
5.0
2.5
0.0
#Reads
P-site position relative to BSJ
[-31;-17]
[-8;6]
Discard mRNA crossmapping
C
Fraction of reads
Upstream (169582)
BSJ-spanning (1578)
Upstream (339001)
BSJ-spanning (1582)
Upstream (921092)
BSJ-spanning (5913)
Upstream (1062824)
BSJ-spanning (8)
D
Fraction
Upstream (118956)
BSJ-spanning (203)
Upstream (209130)
BSJ-spanning (198)
Upstream (509843)
BSJ-spanning (677)
Upstream (617182)
BSJ-spanning (4)
On frame
Off frame(1)
Off frame(2)
E
RSRC1 (34)
UBXN7 (29)
STX6 (11)
AAGAB (14)
CDYL (14)
CFLAR (20)
MTHFD2L (10)
LRCH3 (11)
KIAA0182 (38)
CRIM1 (9)
LDLR (19)
NRIP1 (11)
URI1 (17)
XPO1 (17)
MAN1A2 (115)
XPO1 (17)
ASXL1 (18)
ASXL1 (18)
FKBP8 (13)
C9orf5 (13)
AFF1 (17)

Figure 7


A
circRNA
+ mRNA
circRNA
specific
Host gene
circRNA
decoy
Comet MS/MS search tool
B
−log (E−value)
30
20
10
0
circRNA
Host gene
Decoy
AUG circRNA
Ambiguous AUG circRNA
Other circRNA
AUG exon
C
AUG circRNA
Ambig. AUG circRNA
Other circRNA
AUG exon
Hits+1
1e+05
1e+03
1e+01
0.0
0.1
0.2
0.3
0.4
0.5
FDR cutoff
circRNA
Host gene