HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2021 August 16.

Published in final edited form as:

Cell Rep. 2021 July 27; 36(4): 109439. doi:10.1016/j.celrep.2021.109439.

Mammalian circular RNAs result largely from splicing errors

Chuan Xu[1,2], **Jianzhi Zhang**[2,3,*]

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders of Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China

[2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

[3]Lead contact

SUMMARY

Ubiquitous in eukaryotes, circular RNAs (circRNAs) comprise a large class of mostly non-coding RNAs produced by back-splicing. Although some circRNAs have demonstrated biochemical activities, whether most circRNAs are functional is unknown. Here, we test the hypothesis that circRNA production primarily results from splicing error and so is deleterious instead of beneficial. In support of the error hypothesis, our analysis of RNA sequencing data from 11 shared tissues of humans, macaques, and mice finds that (1) back-splicing is much rarer than linear-splicing, (2) the rate of back-splicing diminishes with the splicing amount, (3) the overall prevalence of back-splicing in a species declines with its effective population size, and (4) circRNAs are overall evolutionarily unconserved. We estimate that more than 97% of the observed circRNA production is deleterious. We identify a small number of functional circRNA candidates, and the genome-wide trend strongly suggests that circRNAs are largely non-functional products of splicing errors.

Graphical Abstract

*Correspondence: jianzhi@umich.edu.

AUTHOR CONTRIBUTIONS

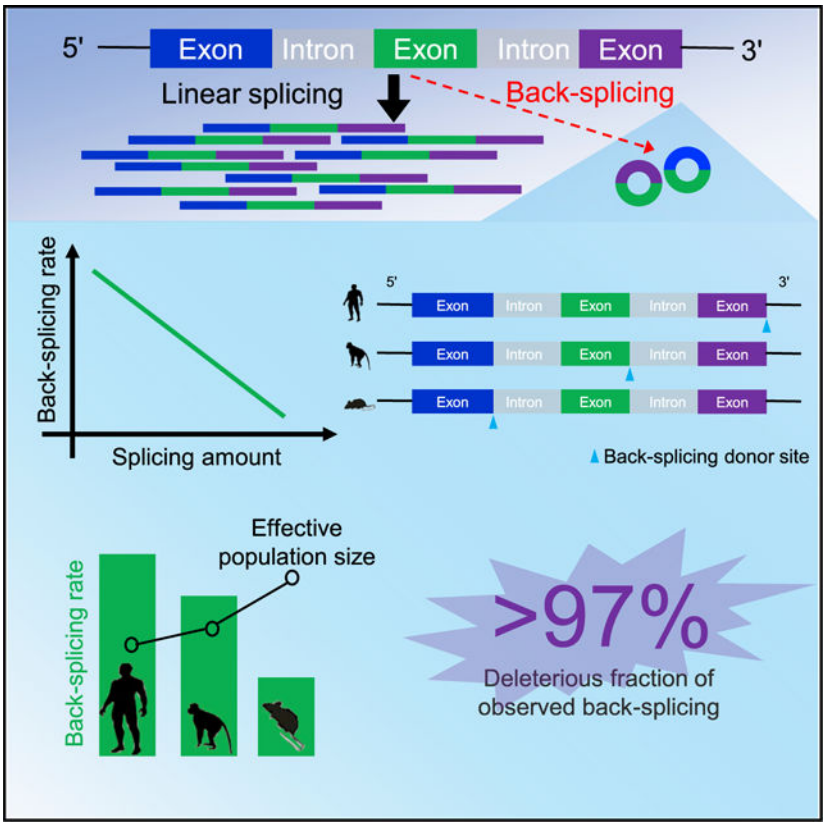
C.X. and J.Z. designed the study and wrote the paper. C.X. performed the study and analyzed the data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109439.



In brief

Circular RNAs (circRNAs) comprise a large class of mostly non-coding RNAs generated through back-splicing of pre-mRNAs, but the biological functions of circRNAs are largely unknown. Xu and Zhang provide evidence that most mammalian circRNAs are products of splicing errors and likely do not confer benefits.

INTRODUCTION

Circular RNAs (circRNAs) are a class of eukaryotic, endogenous, single-stranded, mostly non-coding RNA; unlike the regular RNAs formed by canonical linear-splicing, circRNAs are generated by back-splicing that covalently links a downstream splice-donor site to an upstream splice-acceptor site (Chen, 2016; Kristensen et al., 2019; Vicens and Westhof, 2014). Back-splicing requires canonical splicing signals (Starke et al., 2015), uses the canonical splicing machinery (Kristensen et al., 2019), and competes with canonical pre-mRNA splicing (Ashwal-Fluss et al., 2014). The length and location of circularized exons and the sequence content and length of the flanking introns of the back-spliced sites have been shown to impact circRNA biogenesis (Jeck et al., 2013; Memczak et al., 2013; Salzman et al., 2012; Zhang et al., 2014).

circRNAs generally include canonical exons (Zhang et al., 2016), are predominantly cytoplasmic (Huang et al., 2018; Salzman et al., 2012), and are exceptionally stable (Enuka et al., 2016; Memczak et al., 2013). High-throughput RNA sequencing (RNA-seq) coupled

with circRNA-specific bioinformatics has discovered numerous circRNAs (Glažar et al., 2014; Guo et al., 2014; Ivanov et al., 2015; Jeck et al., 2013; Ji et al., 2019; Salzman et al., 2012; Wang et al., 2014; Westholm et al., 2014). For example, over 50% of human protein-coding genes have been found to produce circRNAs (Ji et al., 2019). circRNAs are specific to tissue (Ji et al., 2019; Xia et al., 2017), cell type (Guo et al., 2014; Salzman et al., 2013), developmental stage (Szabo et al., 2015; Tan et al., 2017), and even subcellular location (Zhang et al., 2019). For instance, many circRNAs are dynamically expressed in the mammalian brain and are enriched in synapses (Ji et al., 2019; Rybak-Wolf et al., 2015; Xia et al., 2017). Some circRNAs act as microRNA sponges (Kristensen et al., 2019; Patop et al., 2019). The best known example is CDR1as/CiRS-7, which carries over 70 binding sites for miR-7, efficiently tethers miR-7, and drastically suppresses miR-7's activity in binding its mRNA targets (Hansen et al., 2013; Memczak et al., 2013). Some circRNAs bind to and titrate out RNA-binding proteins (RBPs) (Abdelmohsen et al., 2017; Ashwal-Fluss et al., 2014). For instance, circMbl, derived from *muscleblind* (MBL/MBNL1), can titrate out extra MBL proteins (Ashwal-Fluss et al., 2014). Additionally, some circRNAs act as scaffolds to mediate the formation of complexes between specific enzymes and substrates (Du et al., 2016) and recruit proteins to particular locations (Chen et al., 2018). Furthermore, a small subset of circRNAs may take effect through their protein products resulting from cap-independent translation (Pamudurti et al., 2017). These demonstrated biochemical activities can be important, although they have been found in only a tiny fraction of all circRNAs. In fact, a genome-wide analysis suggested that most circRNAs are neither microRNA sponges nor translated (Guo et al., 2014).

In the early days after the discovery of circRNAs (Hsu and Coca-Prados, 1979), these molecules were thought to be the product of erroneous splicing (Cocquerelle et al., 1993), a view that we refer to as the error hypothesis of circRNA production. However, the high prevalence of circRNAs, along with the demonstrated biochemical activities of a small number of them, has led to an alternative view that circRNAs are a large group of functional RNAs widely used in gene regulation (Barrett and Salzman, 2016; Chen, 2016; Ebbesen et al., 2017; Kristensen et al., 2019; Li et al., 2018; Memczak et al., 2013; Meng et al., 2017; Patop et al., 2019; Qu et al., 2017; Salzman, 2016). The popularity of this view is reflected by a rapid growth in the interest in circRNAs; only 8 years after the report of circRNAs produced from hundreds of human genes (Salzman et al., 2012), the term circRNAs appeared in the title or abstract of over 2,900 papers in 2020 alone. We will name this now prevailing view the adaptive hypothesis because circRNA production is beneficial according to this view. The adaptive hypothesis includes the scenario of exaptation in which circRNAs originate as functionless molecular errors but have since been co-opted to become functional and beneficial today.

Despite the popularity of the adaptive hypothesis, the error hypothesis is not out of the question for most circRNAs. Back-splicing that creates circRNAs is a type of alternative splicing, which is known to be error prone (Melamud and Moult, 2009; Pickrell et al., 2010; Saudemont et al., 2017). Hence, back-splicing as a splicing error could occur to the transcripts of many genes. Furthermore, the error hypothesis is not inconsistent with the fact that only a tiny fraction of circRNAs have demonstrated biochemical activities. Furthermore, it is unknown how many of these activities are selected and how many have

no appreciable fitness effects (Doolittle et al., 2014; Graur et al., 2013). Distinguishing between the error and adaptive hypotheses of circRNA production is important because it will shed light on the origin, function, and biological significance of this large group of ubiquitous RNAs of eukaryotes and guide future circRNA research. Here, we make a series of distinct predictions of the error hypothesis about genomic patterns of back-splicing and circRNAs that are not expected *a priori* under the adaptive hypothesis. By analyzing high-throughput RNA-seq data from multiple tissues of humans, macaques, and mice, we provide comprehensive evidence that the production of most mammalian circRNAs is due to splicing error and is selectively disfavored.

RESULTS

Back-splicing rates are generally very low

Under the error hypothesis, back-splicing is a splicing error, which is expected to be generally detrimental. Thus, natural selection should have minimized the rate of back-splicing, which is defined as the probability that a splicing event leads to back-splicing instead of linear-splicing. In contrast, the adaptive hypothesis does not predict *a priori* a low rate of back-splicing because, under this hypothesis, back-splicing rates should be high enough to yield sufficient circRNAs for them to have functional impacts (Palazzo and Lee, 2015).

To distinguish between the error and adaptive hypotheses, we investigated back-splicing rates by using a RiboMinus RNA-seq dataset from the human, macaque, and mouse (see STAR Methods). We focused on the 11 tissues in the dataset that are shared among the 3 mammals to facilitate among-species comparisons (Table S1). We then identified linearly spliced reads, which indicate linear-splicing, and back-spliced reads, which indicate back-splicing (see STAR Methods). We define the splicing amount of a gene by the total amount of back-splicing and linear-splicing of the gene. To ensure a certain level of accuracy in the estimation of back-splicing rates, we considered only those protein-coding genes for which the expression level is at least 1 transcript per kilobase million (TPM) and the splicing amount is at least 1 spliced read. As previously reported for this dataset (Ji et al., 2019), a relatively large fraction of genes show back-splicing. Among the 11 tissues, the median fraction of genes exhibiting back-splicing is 27.2%, 37.8%, and 25.5% in the human, macaque, and mouse, respectively (first column in Figure 1A). However, the median fraction of splice sites subject to back-splicing is only 3.9%, 6.2%, and 2.9% for human, macaque, and mouse, respectively (second column in Figure 1A). Most importantly, the median rate of back-splicing, measured by the median fraction of spliced reads that are back-spliced, is only 0.2%, 0.16%, and 0.04% in human, macaque, and mouse, respectively (third column in Figure 1A), indicating that the overall back-splicing rate is three to four orders of magnitude lower than the linear-splicing rate. In any tissue of any of the three species, even when only genes exhibiting back-splicing (i.e., back-spliced genes) in the tissue are considered, back-spliced reads constitute no more than 2% of all spliced reads (fourth column in Figure 1A). We further examined the distribution of the fraction of spliced reads that are back-spliced among back-spliced genes. Again, we found this fraction to be below 10% in human and below 5% in the other species in most genes (Figure 1B). Due to

the exceptional stability of circRNAs relative to linear RNAs, the actual back-splicing rates are likely even lower than the above estimates. Together, these observations show that the back-splicing rate is orders of magnitude lower than the linear-splicing rate, as expected if back-splicing is a splicing error.

Back-splicing rates decrease with splicing amount

Under the error hypothesis, there are at least three reasons why back-splicing is likely detrimental and selected against. First, back-splicing lowers the fraction of functional mRNA molecules. Second, it wastes materials and energy in producing and degrading circRNAs and possibly their protein products. Third, it may result in circRNAs and/or their protein products that are toxic. Under a given rate of back-splicing, the harm of back-splicing due to the above first cause is independent of the total splicing amount but that due to the second and third causes increases with the total amount of splicing. Hence, natural selection against back-splicing at a splice site (or in a gene) should intensify with the amount of splicing at the splice site (or in the gene). As a result, the error hypothesis predicts that the back-splicing rate should decrease with the splicing amount. In contrast, the adaptive hypothesis does not predict this negative correlation *a priori* because, under this hypothesis, the back-splicing rate depends on the specific function and regulation of the gene and/or the circRNA produced.

To distinguish between the error and adaptive hypotheses, for each (expressed and spliced) gene, we estimated its splicing amount by the total number of spliced reads in the gene, which rises with the transcript concentration of the gene as well as its number of introns. Because natural selection against splicing error in a gene depends on the product of the above two variables, we do not consider them separately. We estimated the back-splicing rate of a gene by its proportion of spliced reads that are back-spliced. We started by focusing on back-spliced genes in the human kidney. Consistent with the prediction of the error hypothesis, the rank correlation (ρ) between the splicing amount of a gene and its back-splicing rate is significantly negative ($\rho = -0.63$, $p < 10^{-300}$; Figure 2A). Qualitatively similar results were observed in all examined tissues of the human, macaque, and mouse (Figure 2B). For comparison, we marked in Figure 2A the host genes of two functional circRNAs, namely, circ-ZNF609 (Legnini et al., 2017) and circ-FBXW7 (Yang et al., 2018); the back-splicing rates are much greater in these two genes than in most other genes of comparable splicing amounts.

The above correlation analysis is subject to two potential statistical problems. First, because the detectability of back-splicing increases with the splicing amount, both low and high rates of back-splicing are observable in genes of high splicing amounts whereas only high rates of back-splicing may be observed in genes of low splicing amounts. Consequently, a negative correlation between the back-splicing rate and splicing amount could have resulted simply from this potential detection bias. Second, because the splicing amount is used as the denominator in the estimation of the back-splicing rate, any measurement error of splicing amount can cause a spurious correlation between splicing amount and back-splicing rate. To avoid these potential problems, we used a supergene approach (see STAR Methods). Briefly, we ranked all genes by the splicing amount and grouped them into 10 bins such that each bin

had the same total splicing amount. We then computed the overall back-splicing rate of each bin by considering all genes in the bin together as a supergene. The uniformity of the total splicing amount among bins rids the potential problems mentioned. In the human kidney, the back-splicing rate of a bin decreases almost monotonically with the median splicing amount of all genes in the bin ($\rho = -0.99$, $p < 10^{-300}$; Figure 2C). Similar results were found in all tissues of the three mammals (Figure 2D). Note that this negative correlation cannot be caused by potentially false signals of back-splicing created by sequencing or other technical errors because such errors are random and so should not occur more frequently to genes of lower splicing amounts. Furthermore, the negative correlation could not have been caused by an impact of the amount of back-splicing on the total splicing amount (e.g., under the hypothesis that back-splicing is a functional regulation of gene expression) because the former is such a tiny fraction of the latter (Figure 1) that the variation of the former has effectively no influence on the variation of the latter. To exclude the possibility that the above results are statistical artifacts, we performed a computer simulation analogous to the analysis in Figure 2C, except that we randomly shuffled the back-splicing rates among genes before the analysis. As expected, no significant correlation was observed between the overall back-splicing rate of a bin and the median splicing amount of the genes belonging to the bin.

Different genes differ in multiple aspects in addition to the splicing amount, and so they may not be comparable. To minimize the influences of potential confounding factors in the above analysis, we compared the back-splicing rates between paralogous genes because paralogs are similar in gene structure, DNA sequence, regulation, and function (Zhang, 2013). We required the splicing amount to be at least two times different between the two paralogs to ensure sufficient power of the analysis. Consistent with the error hypothesis, for a paralogous pair, the back-splicing rate tends to be higher for the gene of a relatively low splicing amount. For example, in the human kidney, 75.3% of paralogous pairs show such a trend, which is significantly more than the random expectation of 50% ($p = 1.89 \times 10^{-25}$, binomial test; Figure 2E). In the above analysis, we randomly chose two genes from each gene family annotated by Ensembl. To ensure that the observation in Figure 2E is robust, we repeated the above analysis 100 times; the significant trend in Figure 2E was confirmed in each of the 100 replications. Furthermore, the pattern in Figure 2E holds in all analyzed tissues of the three mammals (Figure 2F). Because the supergene approach cannot pair paralogous genes, we used a downsampling approach (see STAR Methods) to remove the potential statistical problems mentioned. Specifically, for each pair of paralogs, we randomly sampled the spliced reads from the paralog with a relatively high splicing amount to the number of spliced reads observed in the other paralog. Wefoundthe results from the downsampled data to be virtually identical to those from the original data (Figure 2F).

The back-splicing rate correlates negatively with splicing amount across tissues

The back-splicing rate of a gene can be influenced by *cis*-acting elements, which are present on the same DNA molecule as the gene, and *trans*-acting factors, which are not present on the same DNA as the gene. In the above comparison of backing-splicing among different genes in the same tissue, all genes are in the same environment of *trans*-acting factors, so the among-gene variation must be caused by the variation in *cis*-acting elements that

affect splicing. Because different tissues can provide qualitatively or quantitatively different *trans*-acting factors, the back-splicing rate of the same gene may differ among tissues. The error hypothesis predicts that, for a given gene, natural selection against splicing error should intensify in the tissue where the splicing amount of the gene is higher. This should result in a negative correlation between the back-splicing rate and splicing amount across tissues for individual genes. In contrast, no such prediction is made *a priori* by the adaptive hypothesis because, under the adaptive hypothesis, the back-splicing rate of a gene in a tissue would depend on the specific function of the circRNA produced in that tissue.

Because the back-splicing rate is low or even zero for most genes (Figure 1B), sampling error would swamp the potential signal in among-tissue comparisons of individual genes. To circumvent this problem, we randomly grouped every 250 genes into a supergene, except for the last supergene that comprised the remainder of fewer than 250 genes after the grouping. The number 250 was chosen to ensure that each supergene contains sufficient genes with back-splicing and that sufficient supergenes are present to permit a meaningful statistical analysis. We examined the back-splicing rate and splicing amount of each supergene in each tissue. To allow for an among-tissue comparison of the splicing amount, we computed the splicing amount of a supergene in a tissue by the number of spliced reads for the supergene per million total reads (SRPM) in the tissue. As shown in Figure 3A for an example, the back-splicing rate of this particular human supergene in a tissue generally decreases with its splicing amount in the tissue. Indeed, in each of the three species examined, significantly more than 50% of supergenes show this negative correlation (numbers at the bottom of Figure 3B), and this trend is robust as long as supergenes are composed of at least 250 genes. Because each supergene does not have the same splicing amount across tissues, to avoid potential statistical artifacts, we downsampled the spliced reads of a supergene in a tissue to the lowest observed level among all tissues for the supergene and then recomputed the back-splicing rate of the supergene in each tissue. The final results still hold (Figure 3). Thus, the variation of the back-splicing rate among tissues supports the error hypothesis. Because the among-tissue variation of the back-splicing rate is not always concordant among (super) genes, we infer that the variation not only is due to differences in *trans*-acting factors among tissues but also contributed by interactions between *trans*-acting factors of individual tissues and *cis*-acting elements of individual genes.

Back-splicing is not evolutionarily conserved

Back-splicing is expected to be evolutionarily conserved if it is beneficial; otherwise, it should be unconserved. Thus, a comparison between species that have been separated for a sufficiently long time allows differentiation between the adaptive and error hypotheses. To this end, we compared back-splicing between a primate and a rodent. If an orthologous splice-acceptor (or donor) is used in back-splicing in the same tissue of the two species, the acceptor (or donor) is considered shared between the two species (Figure 4A). For each tissue, we calculated the fraction of back-spliced acceptors in human or macaque that are shared with mouse. For example, in the kidney, human has 9,133 back-spliced acceptors, of which only 1,539 (or 16.9%) are shared with mouse. This fraction has a median value of 17.0% in human and 12.3% in macaque across the 11 tissues examined (Figure 4B).

Nevertheless, sharing of a spliced acceptor or donor between species may not indicate functional back-splicing because non-functional back-splicing could be shared by chance. To estimate the amount of sharing expected by chance, we first examined patterns of back-splicing in human (of a given tissue) by computing the relative probabilities that a back-spliced read mapped to a splice donor is also mapped respectively to the first upstream acceptor, second upstream acceptor, and so on (Figure 4A). We then used the overall probability distribution (Data S1) of all back-spliced reads observed in the species and tissue to simulate back-splicing. If n back-spliced reads were observed for a donor, we would simulate n back-spliced reads for this donor, but the acceptors will be randomly decided based on the overall probability distribution determined above. The simulation allows the survey of the set of acceptors expected by chance when the donors are given. We repeated the analysis and surveyed the set of acceptors expected by chance in mouse and then computed the fraction of acceptors in human shared with mouse by chance. For example, this value is 14.6% for the kidney. Hence, the observed fraction of shared acceptors is only 16.9 − 14.6 = 2.3 percentage points above the chance expectation (Figure 4C), indicating that most acceptors shared between human and mouse are by chance. Similar patterns are observed for other human tissues and for all macaque tissues (Figure 4C).

We similarly analyzed the sharing of back-spliced donors between a primate and a rodent (Data S2). Again, we found moderate sharing of donors given the acceptors (Figure 4D) but most are explainable by chance (Figure 4E). Note that some values in Figure 4C and Figure 4E are negative, which is likely due to sampling error caused by the stochasticity of evolution and/or simulation. Regardless, the small positive to small negative values suggest that there is little excess in between-species sharing of back-spliced acceptors (given donors) or donors (given acceptors) when compared with the chance expectation, supporting the error hypothesis and refuting the adaptive hypothesis. Note that some acceptors and donors are used more often than others for back-splicing (Data S1 and S2), but we do not know the molecular determinants of their relative usages, which await future mechanistic studies.

Conservation of splicing signals is uncorrelated with the amount of back-splicing

Back-splicing depends on the canonical splicing machinery and splicing signal. Thus, with proper controls, intraspecific and interspecific variations of splicing motifs—GU as the donor and AG as the acceptor—can indicate whether back-splicing is protected by purifying selection. If back-splicing is functional, motifs associated with larger amounts of back-splicing should be subject to stronger purifying selection. In contrast, if back-splicing results mostly from molecular error and is not beneficial, no such correlation is expected. To this end, we used the number of back-spliced reads associated with a donor (or acceptor) in a human tissue as the measure of its back-splicing amount in that tissue. We used (1) the number of single-nucleotide polymorphisms (SNPs) per site (SNP density) in humans, (2) the mean derived allele frequency (DAF) in humans, and (3) the percent sequence divergence between human and macaque at a donor (or acceptor) splicing motif—the first (or last) two nucleotides of the relevant intron—as indicators of purifying selection. All three indicators should decline with the level of purifying selection but have different properties. The interspecific sequence divergence measures long-term average purifying selection and is insensitive to interferences from selections at linked nucleotide sites but

would be powerless if the functionality of the associated back-splicing is limited to humans. The other two indicators are useful even if the functionality of the associated back-splicing is limited to humans but could be influenced by linked selection. In addition, SNP density could be affected by mutation rate variation among sites, whereas DAF is robust to this variation.

We observed only trivial and mostly statistically non-significant correlations between the back-splicing amount and SNP density among back-spliced donors or acceptors (Figure 5A). Note that these correlations may not be due to selection on back-splicing because the donors and acceptors are also used by linear-splicing. Indeed, we observed significant, negative correlations between the linear-splicing amount and SNP density across donors and acceptors (Figure 5B). To remove the confounding factor of linear-splicing, we performed partial correlations between the back-splicing amount and SNP density by controlling the corresponding liner-splicing amount. Interestingly, all the partial correlations are around zero and none of them are statistically significant (Figure 5C). Similar patterns were observed for DAF (Figures 5D–5F) and interspecific sequence divergence (Figures 5G–5I). Together, both intraspecific polymorphisms and interspecific divergences of donor and acceptor motifs suggest no purifying selection protecting back-splicing motifs, which is inconsistent with the adaptive hypothesis but supports the error hypothesis.

Overall rate of back-splicing declines with the effective population size

If back-splicing arises from splicing error and is detrimental, natural selection will lower its rate. Because the strength of the selection increases with the effective population size (N_e) of the species (Ohta, 1992), the rate of back-splicing upon selection is expected to be lower in species with larger N_e. That is, the error hypothesis predicts that the back-splicing rate reduces from the human to macaque to mouse, given that N_e increases substantially from the human to macaque to mouse (Phifer-Rixey et al., 2012; Xue et al., 2016). In contrast, no such prediction is made *a prior* by the adaptive hypothesis because the back-splicing rate in a species would depend on the function of back-splicing and the environment of the species under the adaptive hypothesis.

To compare the overall rate of back-splicing among the three species, we grouped the splicing data from all 11 tissues of each species. We first calculated the overall back-splicing rate of all (expressed and spliced) genes in each species, which is the total number of back-spliced reads divided by the total number of spliced reads. This rate is 0.26% in humans, 0.19% in macaques, and 0.09% in mice, with all between-species differences being significant ($p < 10^{-15}$, Fisher's exact test; Figure 6A). Because the number and types of genes vary among species, we also compared one-to-one orthologous genes among the three species. Now the back-splicing rate is 0.31% in humans, 0.22% in macaques, and 0.11% in mice (all $p < 10^{-15}$, Fisher's exact test; Figure 6B). We confirmed that the above pattern of interspecific differences holds when orthologous genes are stratified into groups of low (<10 SRPM), intermediate (10–100 SRPM), and high (>100 SRPM) splicing amounts (all $p < 10^{-15}$, Fisher's exact test; Figure 6C). These results are not caused by outliers, which is evident from the among-gene distribution of the back-splicing rate of each species (Figure 6D). Furthermore, a comparison of splicing rates of individual genes among the three

species supports that the splicing rate generally reduces from human to macaque to mouse for orthologous genes (all $p < 10^{-238}$, Wilcoxon signed-rank test; Figure 6D). Together, the relative overall back-splicing rates in the three mammals supports the error hypothesis. Nevertheless, because the above analyses were based on only three species, our finding should be further scrutinized in the future when circRNA data of the same tissues become available from additional species.

Most back-splicing is deleterious

Together, the above analyses strongly suggest that most back-splicing events are deleterious. Below, we use an established method to estimate the fraction of back-splicing that is deleterious (Li and Zhang, 2019; Saudemont et al., 2017; Xu and Zhang, 2020a). This estimation is based on the reasonable assumption that the fitness effect of a back-splicing event at a splice site before the action of natural selection is independent of the splicing amount at the site. Because the strength of natural selection against back-splicing increases with the splicing amount, we assume that all deleterious splicing has been selectively removed in genes of the highest splicing amounts. In other words, the observed back-splicing rate in these genes is the non-deleterious back-splicing rate (ND). Similarly, we assume that none of the deleterious back-splicing has been selectively purged in genes of the lowest splicing amounts. That is, the observed back-splicing rate in these genes reflects the total back-splicing rate (T). Thus, the fraction of deleterious back-splicing is $F_{del} = (T\text{-}ND)/T = 1 - ND/T$. We defined genes of the lowest and highest splicing amounts by using a variety of cutoffs. In theory, using more stringent cutoffs makes the estimate of F_{del} more accurate but less precise due to the reduction in sample size. When the data from all tissues were combined, we found F_{del} to be greater than 96% for each species under any combination of cutoffs (Figure 7A). For example, when the cutoffs of <1 SRPM and >500 SRPM were adopted in defining genes of the lowest and highest splicing amounts, respectively, human $T = 5.96 \times 10^{-3}$ and $ND = 6.46 \times 10^{-5}$, so $F_{del} = 98.9\%$. Similarly, under these cutoffs, $F_{del} = 99.8\%$ for the macaque and 99.1% for the mouse. Note that the above F_{del} values of the three species are not directly comparable because the same SRPM cutoffs mean different degrees of validity of the above two assumptions for different species as a result of their different N_e values.

Under these same cutoffs, we also estimated F_{del} for each tissue in each species. All F_{del} values are >73% except for the human brain, which has an F_{del} of 42% (Figure 7B). Upon examination of the 29 genes of >500 SRPM in the human brain, we found a gene (*RIMS1*) with an unusually high back-splicing rate of 8.7%. Because the circRNAs produced from *RIMS1* were reported to be potentially functional in neurons (Chen et al., 2019; Ji et al., 2019; You et al., 2015), we re-estimated F_{del} after removing *RIMS1*. Now, $F_{del} = 91.6\%$ in the human brain and remains virtually unchanged in the other tissues or species (Figure 7B). Although F_{del} varies among tissues, we observed a median F_{del} of 98.8%, 98.4%, and 98.0% in the human, macaque, and mouse, respectively (Figure 7B), which are similar to the F_{del} estimates from all tissues together (Figure 7A). Note that our F_{del} estimates are conservative because very slightly deleterious back-splicing may not have been fully removed by selection in the genes of the highest splicing amounts and because some

strongly deleterious back-splicing may have been removed by selection even in the genes of the lowest splicing amounts.

Note that F_{del} measures the fraction of back-splicing that is deleterious before the action of purifying selection. Because some deleterious back-splicing has been removed by selection, the fraction of observed back-splicing that is deleterious should be lower. The deleterious fraction of observed back-splicing (O_{del}) can be estimated by regarding the overall back-splicing rate from all genes (Figure 6A) as T and the back-splicing rate from genes of >500 SRPM as ND. When the data from all tissues are merged, O_{del} is 97.5% for human, 99.7% for macaque, and 98.9% for mouse, respectively. The O_{del} values are only slightly lower than the corresponding F_{del} values because apparently most deleterious back-splicing has not been selectively purged due to the preponderance of genes of relatively low splicing amounts. It has been estimated that the fraction of linear-splicing that arises from error is about 70% in humans (Saudemont et al., 2017), suggesting that error accounts for a much greater proportion of back-splicing than linear-splicing. Taken together, our estimation demonstrates that the vast majority of all or observed back-splicing is deleterious, which is broadly consistent with the finding of virtually no excess in between-species sharing of back-spliced acceptors or donors over the chance expectation and the finding of no purifying selection protecting back-splicing signals.

DISCUSSION

The discovery of a large number of circRNAs from many eukaryotes and the demonstration that some of them possess biochemical activities have led to the prevailing view that circRNA production is generally beneficial. In this work, we challenged this adaptive view by providing comprehensive evidence for an alternative view that back-splicing that leads to circRNA production arises mostly from splicing error and is detrimental. Our evidence, based on the transcriptomes of 11 tissues from each of the human, macaque, and mouse, comprises the findings that (1) the back-splicing rate is orders of magnitude lower than the linear-splicing rate, (2) the back-splicing rate in a gene decreases with the splicing amount of the gene, (3) the back-splicing rate of a gene in a tissue tends to reduce with the splicing amount of the gene in the tissue when multiple tissues are compared, (4) there is little between-species sharing of back-spliced acceptors or donors beyond the chance expectation, (5) purifying selection protecting the motifs for back-splicing is lacking, and (6) the overall rate of back-splicing in a species declines with its effective population size. None of these observations are predicted *a prior* by the adaptive hypothesis, but all fit the predictions of the error hypothesis. Although most of the evidence is derived from the 11 tissues analyzed, the above fifth line of evidence is based on the polymorphism and divergence of genome sequences and so is not limited to the specific tissues analyzed. Together, the empirical evidence strongly suggests that most mammalian back-splicing events and by inference most circRNA productions are detrimental rather than beneficial.

Aside from the above evidence, there are several observations reported in the literature that are consistent with the error hypothesis or inconsistent with the adaptive hypothesis. First, the proposal that binding to microRNAs is a general function of circRNAs has been challenged (Enuka et al., 2016; Guo et al., 2014; Ragan et al., 2019). For example, Ragan et

al. (2019) reported that only about 12% of circRNAs contain microRNA binding sites. Guo et al. (2014) found only two circRNAs with more microRNA binding sites than expected by chance, and Enuka et al. (2016) found no enrichment of microRNA binding sites in circRNAs in general. Second, although some circRNAs can act as the sponge of RBPs (Zang et al., 2020), this activity is not generalizable because a genome-wide analysis did not find enriched binding sites of RBPs in circRNAs compared with their corresponding linear mRNAs (You et al., 2015). Third, researchers failed to detect a significant association of circRNAs with polysomes (Guo et al., 2014; You et al., 2015), arguing against the notion that circRNAs generally function through their protein products (Kristensen et al., 2019; Li et al., 2018; Pamudurti et al., 2017). circRNA translation depends on having internal ribosome entry sites (IRESs) (Pamudurti et al., 2017), but <1.5% of circRNAs contain IRESs (Fan et al., 2019). Recently, however, IRES-like elements were found in many circRNAs, and hundreds of circRNA-encoded peptides were identified from mass spectrometry data (Fan et al., 2019). Notwithstanding, the translation of a circRNA does not prove that it confers an advantage because the translation itself could be an error due to spurious translational initiation and the protein product could be functionless or even toxic. In fact, the sequence similarity of circRNA orthologs between human and mouse is no higher than that of their neighboring linear exons, suggesting a lack of circRNA-specific purifying selection (Guo et al., 2014). Finally, it is worth stressing that the existence of regulation or regulatory mechanisms of circRNA biogenesis or degradation (Conn et al., 2015; Liang et al., 2017; Zhang et al., 2014) is not evidence for circRNA functionality because this phenomenon can arise as a byproduct of other biological processes. As an analogy, simply because trash is removed once a week on a particular day, the amount of trash in a house would exhibit a cyclic pattern resembling regulation, but the regulation does not prove that the trash is useful.

It is worth noting that the brain shows more back-spliced genes and a higher back-splicing rate than the other 10 tissues studied here (Figure 1). Nevertheless, for the following four reasons, this observation does not necessarily support the proposal that circRNAs play important roles in the brain (Rybak-Wolf et al., 2015). First, even in the brain, the back-splicing rate is still very low (Figure 1), and other analyses (Figures 2, 3, 4, 5, and 6) do not suggest that the brain is an outlier of the general patterns of circRNAs observed across tissues. Second, there is no between-species sharing of back-spliced acceptors or donors beyond the chance expectation in the brain (Figure 4C and 4E), and the fraction of deleterious back-splicing is at least 73.2% (Figure 7B) in the brain. Third, alternative splicing and its regulation are more complex in the brain than in other tissues (Raj and Blencowe, 2015), which might increase the chance of splicing error. Fourth, because brain cells typically have longer lifespans than other cells (Magrassi et al., 2013), circRNAs, which are exceptionally stable, might accumulate to a higher level in the brain than in other tissues, explaining why the back-splicing rate looks higher (Figure 1A) and the fraction of deleterious back-splicing (Figure 7B) looks lower in the brain than in other tissues.

Taken together, our findings and those discussed above provide unequivocal evidence that most back-splicing events are detrimental splicing errors and that most circRNAs do not have beneficial functions. Thus, circRNA is generally a class of junk RNA (Brosius, 2005; Palazzo and Lee, 2015). This conclusion requires a paradigm shift in circRNA research.

Instead of assuming that all or most circRNAs are functional in today's research, our conclusion requires that we treat all circRNAs as non-functional until proven otherwise. Nevertheless, our conclusion does not preclude the occasional observation of circRNAs that possess beneficial functions, as has been suggested for CDR1as/CiRS-7 and several other circRNAs mentioned in the Introduction. In this context, it is highly valuable to identify functional circRNAs even though they account for only a small fraction of all circRNAs. Although identifying functional circRNAs is beyond the scope of the present study, we explored the possibility that highly expressed genes with exceptionally high back-splicing rates host functional circRNAs. Specifically, we regressed between the back-splicing rate and splicing amount across genes in Figure 2A and calculated Cook's distance for each gene (see STAR Methods). We then defined a gene as an outlier if its Cook's distance is more than 4 times the mean Cook's distance of all genes. Among these outliers, 15 genes have a back-splicing rate of at least 10% and an expression level of at least 10 TPM (Table S2). Interestingly, these 15 genes include the host genes of 2 known functional circRNAs aforementioned, namely, circ-ZNF609 and circ-FBXW7 (Figure 2A). We suggest that the circRNAs from the remaining 13 genes be studied in the future as candidates of functional circRNAs and the approach proposed above for identifying potentially functional circRNAs be systematically evaluated.

Our findings on back-splicing, along with previous studies on linear-splicing (Melamud and Moult, 2009; Pickrell et al., 2010; Saudemont et al., 2017), demonstrate that splicing is generally error prone. Together, they echo other recent findings that a number of steps in transcription and translation are fallible, including, for example, transcriptional initiation, RNA synthesis, polyadenylation, posttranscriptional modification, translational initiation, translational elongation or decoding, translational termination, and posttranslational modification (Gout et al., 2017; Landry et al., 2009; Li and Zhang, 2019; Liu and Zhang, 2018a, 2018b; Park and Zhang, 2011; Ribas de Pouplana et al., 2014; Xu et al., 2019; Xu and Zhang, 2014, 2018, 2020b). These findings indicate that cellular life is far less perfected than is commonly portrayed, which has broad and profound implications for biology (Lynch, 2007, 2014; Warnecke and Hurst, 2011; Zhang and Yang, 2015).

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jianzhi Zhang (jianzhi@umich.edu).

Materials availability—This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the Key resources table.
- This paper does not report original code.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The original datasets analyzed in the present study are provided in the Key resources table.

METHOD DETAILS

Linear-splicing and back-splicing—Because the regular RNA-seq captures poly(A)-enriched linear RNAs, many other RNA species that are not linear or do not have poly(A) tails are lost. RNase R treatment during the RNA library construction can efficiently enrich back-spliced circRNAs but will filter out linear RNAs. Thus, neither type of datasets is appropriate for our study. To identify back-splicing and linear-splicing simultaneously, we chose data from RNA-seq experiments that only remove rRNAs by RiboMinus treatment during the library construction. The recently published RNA-seq data by Ji et al. (2019), comprising deeply sequenced transcriptomes from 11 tissues each from humans, macaques, and mice (Table S1), fulfill our requirements. We downloaded the data from NGDC (https://ngdc.cncb.ac.cn/).

Back-splicing is inferred from RNA-seq reads that span back-spliced junctions and therefore map non-linearly to the genome. Several pipelines and algorithms have been developed to identify specifically non-linear reads and predict the landscape of circRNAs (Szabo and Salzman, 2016). According to comparisons of many circRNA prediction pipelines (Hansen et al., 2016; Zeng et al., 2017), we chose CIRCexplorer2 (Zhang et al., 2016) and CIRI2 (Gao et al., 2018) to infer back-splicing. The former tool uses STAR (Dobin et al., 2013) as the mapper and is dependent on gene annotations, while the latter is based on the mapper BWA (Li and Durbin, 2009) and predicts back-splicing *de novo*. We first used CIRCexplorer2 under default parameters to identify back-spliced sites and associated back-spliced reads. We then added back-spliced sites identified by CIRI2 under default parameters that were not reported by CIRCexplorer2 and added the associated back-spliced reads.

Linear-splicing was retrieved by Tophat2 (Kim et al., 2013) using default parameters, including annotated and newly identified splicing junctions. The genome assemblies used were GRCh38 for human, Mmul 8.0.1 for macaque, and GRCm38 for mouse, all downloaded with gene annotations from Ensembl release 89. Although the genomic annotation is less extensive for the macaque than for the human and mouse, this variation should not bias our analyses because (1) we used *de novo* identification of back- and linear-splicing sites in addition to annotations and (2) most of our analyses were within-species comparisons.

Because the start and end of a gene are variable due to alternative transcriptional initiation and alternative polyadenylation, we defined a gene from the nucleotide that is 500 bp upstream the annotated 5′-most transcriptional start site (Forrest et al., 2014) to the nucleotide that is 1000 bp downstream the annotated 3′-most polyadenylation site (Derti et al., 2012). Any splicing junction located in a defined gene region was considered to

belong to the gene. Only splicing junctions uniquely mapped to a gene were considered in the study.

Splicing amount—The total number of linear- and back-spliced reads mapped to a splice site is the splicing amount of the splice site. The total number of linear- and back-spliced reads mapped to all splice sites of a gene is the splicing amount of the gene. To allow comparing the splicing amount of a gene among samples, we computed the number of *s*pliced *r*eads *p*er *m*illion total reads in the sample (SRPM). The back-splicing rate at a splice site, in a gene, or in a supergene was calculated as the number of back-spliced reads mapped to the site, gene, or supergene, relative to the total number of spliced reads mapped to the site, gene, or supergene.

We used RNA-seq downloaded from NGDC to measure gene expression levels (Table S1). The reads were mapped to the human (GRCh38), macaque (Mmul 8.0.1), or mouse (GRCm38) genome using TopHat2 (Kim et al., 2013). *F*ragment *p*er *k*ilobase of transcripts per *m*illion mapped reads (FPKM) of a gene was first calculated by cufflinks (Trapnell et al., 2012) and then converted to TPM using the formula of TPM = (FPKM × 10^6)/(sum of FPKM) (Li and Dewey, 2011). Only genes expressed (TPM ≥ 1) and spliced (# of spliced reads ≥ 1) were considered in our study.

Corrections of unequal surveys of splicing events among genes—Due to the variation of the splicing amount among genes, splicing is surveyed more extensively for some genes than other genes by RiboMinus RNA-seq. To remove the potential influence of this unequal survey, we used two different approaches unless otherwise mentioned. The first is the supergene approach. Unless otherwise noted, we first ranked all genes by their splicing amounts. We then grouped the genes into 10 bins representing 10 supergenes, requiring the total splicing amount per bin to be the same for all bins. Numbers of back-spliced, linear-spliced, and spliced reads were respectively summed up across all genes in the bin. The supergene approach cannot be used under certain circumstances. Under these circumstances, we used downsampling. For example, when comparing a pair of paralogous genes in a sample, we randomly picked the number of spliced reads from the gene of the relatively high splicing amount to the level observed in the gene of the relatively low splicing amount. This downsampling equalized the survey depth of splicing between the two genes. The supergene approach is preferred over downsampling when both can be used, because the former uses all data while the latter uses only part of the data. In the analysis among 11 tissues, because many back-spliced genes would have no back-spliced reads in multiple tissues upon downsampling, causing a loss of statistical power, we combined the supergene approach with downsampling, as described in Results.

Cook's distance—Cook's distance (D_i) is used in regression analysis to find influential outliers in a set of predictor variables (Cook, 1977). An observation with Cook's distance larger than four times the mean Cook's distance was deemed an outlier in this study. Cook's distance of gene i is

$$D_i = \frac{\sum_{j=1}^{n} \left[\hat{y}_j - \hat{y}_{j(i)}\right]^2}{p \cdot MSE},$$

where $\hat{y}_j$ is the jth gene's fitted response value, $\hat{y}_{j(i)}$ is the jth gene's fitted response value when gene i is removed, n is the number of genes, MSE is the mean squared error, and p is the number of coefficients in the regression model.

Paralogs and orthologs—Paralogous genes were downloaded from Ensembl (release 89; May 2017) for the three species. We obtained 3,678 human protein-coding gene families with 51,657 pairs of paralogs, 3,912 macaque gene families with 46,718 pairs of paralogs, and 3,856 mouse gene families with 79,968 pairs of paralogs, respectively. We randomly selected from each gene family only one paralogous pair that exhibits a two-fold or greater difference in splicing amount to allow a sufficient statistical power.

Orthologous genes among human, macaque, and mouse were downloaded from Ensembl (release 89; May 2017), and only one-to-one orthologous genes were considered in our analysis. The numbers of orthologs between human and macaque, between human and mouse, and between macaque and mouse are 19,754, 16,797, and 15,170, respectively. From these data, we obtained 14,882 one-to-one orthologs among the three mammals. To identify orthologous splice sites between species, we used the UCSC liftOver tool (https://genome.ucsc.edu/util.html) to align the splice sites of one species with the genome of another species.

Polymorphism and divergence—Human polymorphism data, including allele frequencies, from Interim Phase 3 of the 1000 Genomes project (Sudmant et al., 2015), were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/ (last accessed Feb. 28, 2020). This dataset comprises the genotypes of 2,504 individuals from 26 populations and includes a total of 78,136,341 autosomal SNPs. Only SNPs were included in the analysis. The nucleotide observed at a SNP was categorized as ancestral if it is the same as the nucleotide of the 'AA' field in the polymorphism VCF file; other nucleotides at the SNP are derived. The derived allele frequency at a SNP is the frequency of the derived allele at the SNP. Nucleotide differences at splice-acceptors and donors were based on a comparison between human (GRCh38) and macaque (Mmul 8.0.1) genomes through liftOver from UCSC.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses are described in Results, figure legends, and the above Method details section. R was used in statistical analysis (see Key resources table).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Dr. Fangqing Zhao for sharing the data from Ji et al. (2019) and members of the Zhang laboratory for valuable comments. This work was supported in part by the U.S. National Institutes of Health research grant R35GM139484 to J.Z.

REFERENCES

Abdelmohsen K, Panda AC, Munk R, Grammatikakis I, Dudekula DB, De S, Kim J, Noh JH, Kim KM, Martindale JL, and Gorospe M (2017). Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1. RNA Biol. 14, 361–369. [PubMed: 28080204]

Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, and Kadener S (2014). circRNA biogenesis competes with pre-mRNA splicing. Mol. Cell 56, 55–66. [PubMed: 25242144]

Barrett SP, and Salzman J (2016). Circular RNAs: analysis, expression and potential functions. Development 143, 1838–1847. [PubMed: 27246710]

Brosius J (2005). Waste not, want not—transcript excess in multicellular eukaryotes. Trends Genet. 21, 287–288. [PubMed: 15851065]

Chen LL (2016). The biogenesis and emerging roles of circular RNAs. Nat. Rev. Mol. Cell Biol 17, 205–211. [PubMed: 26908011]

Chen N, Zhao G, Yan X, Lv Z, Yin H, Zhang S, Song W, Li X, Li L, Du Z, et al. (2018). A novel FLI1 exonic circular RNA promotes metastasis in breast cancer by coordinately regulating TET1 and DNMT1. Genome Biol. 19, 218. [PubMed: 30537986]

Chen BJ, Huang S, and Janitz M (2019). Changes in circular RNA expression patterns during human foetal brain development. Genomics 111, 753–758. [PubMed: 29709512]

Cocquerelle C, Mascrez B, Hétuin D, and Bailleul B (1993). Mis-splicing yields circular RNA molecules. FASEB J. 7, 155–160. [PubMed: 7678559]

Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, Roslan S, Schreiber AW, Gregory PA, and Goodall GJ (2015). The RNA binding protein quaking regulates formation of circRNAs. Cell 160, 1125–1134. [PubMed: 25768908]

Cook RD (1977). Detection of influential observation in linear regression. Technometrics 19, 15–18.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, and Babak T (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. 22, 1173–1183. [PubMed: 22454233]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. [PubMed: 23104886]

Doolittle WF, Brunet TD, Linquist S, and Gregory TR (2014). Distinguishing between “function” and “effect” in genome biology. Genome Biol. Evol 6, 1234–1237. [PubMed: 24814287]

Du WW, Yang W, Liu E, Yang Z, Dhaliwal P, and Yang BB (2016). Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. Nucleic Acids Res. 44, 2846–2858. [PubMed: 26861625]

Ebbesen KK, Hansen TB, and Kjems J (2017). Insights into circular RNA biology. RNA Biol. 14, 1035–1045. [PubMed: 27982727]

Enuka Y, Lauriola M, Feldman ME, Sas-Chen A, Ulitsky I, and Yarden Y (2016). Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. Nucleic Acids Res. 44, 1370–1383. [PubMed: 26657629]

Fan X, Yang Y, and Wang Z (2019). Pervasive translation of circular RNAs driven by short IRES-like elements. bioRxiv. 10.1101/473207.

Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, et al.; FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level mammalian expression atlas. Nature 507, 462–470. [PubMed: 24670764]

Gao Y, Zhang J, and Zhao F (2018). Circular RNA identification based on multiple seed matching. Brief. Bioinform 19, 803–810. [PubMed: 28334140]

Glažar P, Papavasileiou P, and Rajewsky N (2014). circBase: a database for circular RNAs. RNA 20, 1666–1670. [PubMed: 25234927]

Gout JF, Li W, Fritsch C, Li A, Haroon S, Singh L, Hua D, Fazelinia H, Smith Z, Seeholzer S, et al. (2017). The landscape of transcription errors in eukaryotic cells. Sci. Adv 3, e1701484. [PubMed: 29062891]

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, and Elhaik E (2013). On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol 5, 578–590. [PubMed: 23431001]

Guo JU, Agarwal V, Guo H, and Bartel DP (2014). Expanded identification and characterization of mammalian circular RNAs. Genome Biol. 15, 409. [PubMed: 25070500]

Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, and Kjems J (2013). Natural RNA circles function as efficient microRNA sponges. Nature 495, 384–388. [PubMed: 23446346]

Hansen TB, Venø MT, Damgaard CK, and Kjems J (2016). Comparison of circular RNA prediction tools. Nucleic Acids Res. 44, e58. [PubMed: 26657634]

Hsu MT, and Coca-Prados M (1979). Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature 280, 339–340. [PubMed: 460409]

Huang C, Liang D, Tatomer DC, and Wilusz JE (2018). A length-dependent evolutionarily conserved pathway controls nuclear export of circular RNAs. Genes Dev. 32, 639–644. [PubMed: 29773557]

Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C, and Rajewsky N (2015). Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. Cell Rep. 10, 170–177. [PubMed: 25558066]

Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, and Sharpless NE (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA 19, 141–157. [PubMed: 23249747]

Ji P, Wu W, Chen S, Zheng Y, Zhou L, Zhang J, Cheng H, Yan J, Zhang S, Yang P, and Zhao F (2019). Expanded expression landscape and prioritization of circular RNAs in mammals. Cell Rep. 26, 3444–3460.e5. [PubMed: 30893614]

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36. [PubMed: 23618408]

Kristensen LS, Andersen MS, Stagsted LVW, Ebbesen KK, Hansen TB, and Kjems J (2019). The biogenesis, biology and characterization of circular RNAs. Nat. Rev. Genet 20, 675–691. [PubMed: 31395983]

Landry CR, Levy ED, and Michnick SW (2009). Weak functional constraints on phosphoproteomes. Trends Genet. 25, 193–197. [PubMed: 19349092]

Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, Fatica A, Santini T, Andronache A, Wade M, et al. (2017). Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. Mol. Cell 66, 22–37.e9. [PubMed: 28344082]

Li B, and Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. [PubMed: 21816040]

Li H, and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. [PubMed: 19451168]

Li C, and Zhang J (2019). Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. PLoS Genet. 15, e1008141. [PubMed: 31120886]

Li X, Yang L, and Chen LL (2018). The biogenesis, functions, and challenges of circular RNAs. Mol. Cell 71, 428–442. [PubMed: 30057200]

Liang D, Tatomer DC, Luo Z, Wu H, Yang L, Chen LL, Cherry S, and Wilusz JE (2017). The output of protein-coding genes shifts to circular RNAs when the pre-mRNA processing machinery is limiting. Mol. Cell 68, 940–954.e3. [PubMed: 29174924]

Liu Z, and Zhang J (2018a). Human C-to-U coding RNA editing is largely nonadaptive. Mol. Biol. Evol 35, 963–969. [PubMed: 29385526]

Liu Z, and Zhang J (2018b). Most m6A RNA modifications in protein-coding regions are evolutionarily unconserved and likely nonfunctional. Mol. Biol. Evol 35, 666–675. [PubMed: 29228327]

Lynch M (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. USA 104, 8597–8604. [PubMed: 17494740]

Lynch M, Field MC, Goodson HV, Malik HS, Pereira-Leal JB, Roos DS, Turkewitz AP, and Sazer S (2014). Evolutionary cell biology: two origins, one objective. Proc. Natl. Acad. Sci. USA 111, 16990–16994. [PubMed: 25404324]

Magrassi L, Leto K, and Rossi F (2013). Lifespan of neurons is uncoupled from organismal lifespan. Proc. Natl. Acad. Sci. USA 110, 4374–4379. [PubMed: 23440189]

Melamud E, and Moult J (2009). Stochastic noise in splicing machinery. Nucleic Acids Res. 37, 4873–4886. [PubMed: 19546110]

Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 495, 333–338. [PubMed: 23446348]

Meng X, Li X, Zhang P, Wang J, Zhou Y, and Chen M (2017). Circular RNA: an emerging key player in RNA world. Brief. Bioinform 18, 547–557. [PubMed: 27255916]

Ohta T (1992). The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst 23, 263–286.

Palazzo AF, and Lee ES (2015). Non-coding RNA: what is functional and what is junk? Front. Genet 6, 2. [PubMed: 25674102]

Pamudurti NR, Bartok O, Jens M, Ashwal-Fluss R, Stottmeister C, Ruhe L, Hanan M, Wyler E, Perez-Hernandez D, Ramberger E, et al. (2017). Translation of CircRNAs. Mol. Cell 66, 9–21.e7. [PubMed: 28344080]

Park C, and Zhang J (2011). Genome-wide evolutionary conservation of N-glycosylation sites. Mol. Biol. Evol 28, 2351–2357. [PubMed: 21355035]

Patop IL, Wüst S, and Kadener S (2019). Past, present, and future of circRNAs. EMBO J. 38, e100836. [PubMed: 31343080]

Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK, and Nachman MW (2012). Adaptive evolution and effective population size in wild house mice. Mol. Biol. Evol 29, 2949–2955. [PubMed: 22490822]

Pickrell JK, Pai AA, Gilad Y, and Pritchard JK (2010). Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. 6, e1001236. [PubMed: 21151575]

Qu S, Zhong Y, Shang R, Zhang X, Song W, Kjems J, and Li H (2017). The emerging landscape of circular RNA in life processes. RNA Biol. 14, 992–999. [PubMed: 27617908]

Ragan C, Goodall GJ, Shirokikh NE, and Preiss T (2019). Insights into the biogenesis and potential functions of exonic circular RNA. Sci. Rep 9, 2048. [PubMed: 30765711]

Raj B, and Blencowe BJ (2015). Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles. Neuron 87, 14–27. [PubMed: 26139367]

Ribas de Pouplana L, Santos MA, Zhu JH, Farabaugh PJ, and Javid B (2014). Protein mistranslation: friend or foe? Trends Biochem. Sci 39, 355–362. [PubMed: 25023410]

Rybak-Wolf A, Stottmeister C, Gla zar P, Jens M, Pino N, Giusti S, Hanan M, Behm M, Bartok O, Ashwal-Fluss R, et al. (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol. Cell 58, 870–885. [PubMed: 25921068]

Salzman J (2016). Circular RNA expression: its potential regulation and function. Trends Genet. 32, 309–316. [PubMed: 27050930]

Salzman J, Gawad C, Wang PL, Lacayo N, and Brown PO (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS One 7, e30733. [PubMed: 22319583]

Salzman J, Chen RE, Olsen MN, Wang PL, and Brown PO (2013). Cell-type specific features of circular RNA expression. PLoS Genet. 9, e1003777. [PubMed: 24039610]

Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, Meyer E, and Duret L (2017). The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol. 18, 208. [PubMed: 29084568]

Starke S, Jost I, Rossbach O, Schneider T, Schreiner S, Hung LH, and Bindereif A (2015). Exon circularization requires canonical splice signals. Cell Rep. 10, 103–111. [PubMed: 25543144]

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al.; 1000 Genomes Project Consortium (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81. [PubMed: 26432246]

Szabo L, and Salzman J (2016). Detecting circular RNAs: bioinformatic and experimental challenges. Nat. Rev. Genet 17, 679–692. [PubMed: 27739534]

Szabo L, Morey R, Palpant NJ, Wang PL, Afari N, Jiang C, Parast MM, Murry CE, Laurent LC, and Salzman J (2015). Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol. 16, 126. [PubMed: 26076956]

Tan WL, Lim BT, Anene-Nzelu CG, Ackers-Johnson M, Dashi A, See K, Tiang Z, Lee DP, Chua WW, Luu TD, et al. (2017). A landscape of circular RNA expression in the human heart. Cardiovasc. Res 113, 298–309. [PubMed: 28082450]

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc 7, 562–578. [PubMed: 22383036]

Vicens Q, and Westhof E (2014). Biogenesis of Circular RNAs. Cell 159, 13–14. [PubMed: 25259915]

Wang PL, Bao Y, Yee MC, Barrett SP, Hogan GJ, Olsen MN, Dinneny JR, Brown PO, and Salzman J (2014). Circular RNA is expressed across the eukaryotic tree of life. PLoS One 9, e90859. [PubMed: 24609083]

Warnecke T, and Hurst LD (2011). Error prevention and mitigation as forces in the evolution of genes and genomes. Nat. Rev. Genet 12, 875–881. [PubMed: 22094950]

Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, and Lai EC (2014). Genome-wide analysis of Drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell Rep. 9, 1966–1980. [PubMed: 25544350]

Xia S, Feng J, Lei L, Hu J, Xia L, Wang J, Xiang Y, Liu L, Zhong S, Han L, and He C (2017). Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. Brief. Bioinform 18, 984–992. [PubMed: 27543790]

Xu G, and Zhang J (2014). Human coding RNA editing is generally nonadaptive. Proc. Natl. Acad. Sci. USA 111, 3769–3774. [PubMed: 24567376]

Xu C, and Zhang J (2018). Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. Cell Syst. 6, 734–742.e4. [PubMed: 29886108]

Xu C, and Zhang J (2020a). A different perspective on alternative cleavage and polyadenylation. Nat. Rev. Genet 21, 63.

Xu C, and Zhang J (2020b). Mammalian alternative translation initiation is mostly nonadaptive. Mol. Biol. Evol 37, 2015–2028. [PubMed: 32145028]

Xu C, Park JK, and Zhang J (2019). Evidence that alternative transcriptional initiation is largely nonadaptive. PLoS Biol. 17, e3000197. [PubMed: 30883542]

Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, White S, Dahdouli M, Rio Deiros D, Below JE, Salerno W, et al. (2016). The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. Genome Res. 26, 1651–1662. [PubMed: 27934697]

Yang Y, Gao X, Zhang M, Yan S, Sun C, Xiao F, Huang N, Yang X, Zhao K, Zhou H, et al. (2018). Novel role of FBXW7 circular RNA in repressing glioma tumorigenesis. J. Natl. Cancer Inst 110, 304–315.

You X, Vlatkovic I, Babic A, Will T, Epstein I, Tushev G, Akbalik G, Wang M, Glock C, Quedenau C, et al. (2015). Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. Nat. Neurosci 18, 603–610. [PubMed: 25714049]

Zang J, Lu D, and Xu A (2020). The interaction of circRNAs and RNA binding proteins: An important part of circRNA maintenance and function. J. Neurosci. Res 98, 87–97. [PubMed: 30575990]

Zeng X, Lin W, Guo M, and Zou Q (2017). A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput. Biol 13, e1005420. [PubMed: 28594838]

Zhang J (2013). Gene duplication. In The Princeton Guide to Evolution, Losos J, ed. (Princeton University Press), pp. 397–405.

Zhang J, and Yang JR (2015). Determinants of the rate of protein sequence evolution. Nat. Rev. Genet 16, 409–420. [PubMed: 26055156]

Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, and Yang L (2014). Complementary sequence-mediated exon circularization. Cell 159, 134–147. [PubMed: 25242744]

Zhang XO, Dong R, Zhang Y, Zhang JL, Luo Z, Zhang J, Chen LL, and Yang L (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. 26, 1277–1287. [PubMed: 27365365]

Zhang J, Zhang X, Li C, Yue L, Ding N, Riordan T, Yang L, Li Y, Jen C, Lin S, et al. (2019). Circular RNA profiling provides insights into their subcellular distribution and molecular characteristics in HepG2 cells. RNA Biol. 16, 220–232. [PubMed: 30614753]

Highlights

- The rate of back-splicing for a gene declines with its degree of splicing
- The abundance of back-splicing in a species declines with its effective population size
- Mammalian circRNAs are overall evolutionarily non-conserved
- More than 97% of the observed circRNA production is deleterious

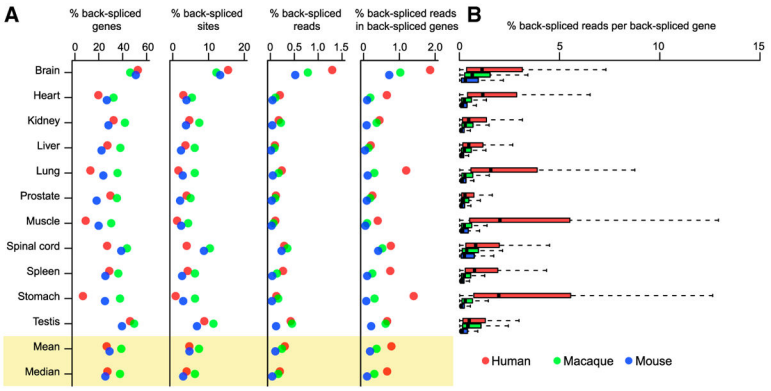


Figure 1. Low rates of back-splicing in mammals, see also Table S1

(A) Various measures of the rate of back-splicing in 11 tissues from 3 mammals. Only expressed and spliced genes are considered. From the left to the right are percentage of genes with back-splicing, percentage of splice sites that show back-splicing, percentage of spliced reads that are back-spliced, and percentage of spliced reads that are back-spliced among back-spliced genes.

(B) Distribution of the percentage of spliced reads that are back-spliced among back-spliced genes. In each boxplot, the left and right edges of a box represent the first (qu1) and third (qu3) quartiles, respectively; the vertical line inside the box indicates the median (md); and the whiskers extend to the most extreme values inside inner fences, md ± 1.5(qu3 − qu1).

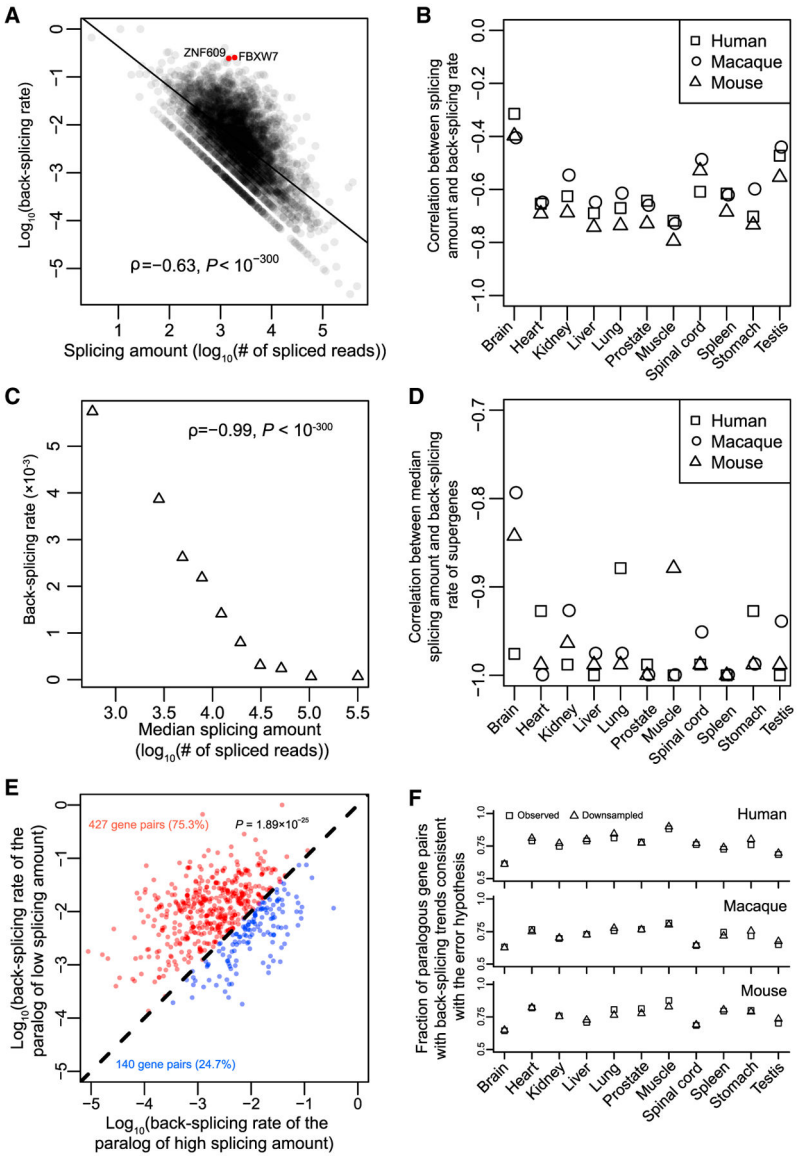


Figure 2. The back-splicing rate of a gene (or supergene) decreases with the splicing amount of the gene (or supergene)

The rate of back-splicing is measured by the fraction of spliced reads that are back-spliced, see also Table S2.

(A) The back-splicing rate of a gene decreases with the splicing amount of the gene in the human kidney. Each dot represents a back-spliced gene, and the solid line shows the linear least-squares regression. Spearman's rank correlation (ρ) and associated p value are presented. The host genes of two functional circRNAs are marked in red.

(B) Spearman's correlation between the back-splicing rate of a gene and its splicing amount among back-spliced genes in each tissue of each mammal examined. All correlations have $p < 10^{-126}$.

(C) The back-splicing rate of a supergene decreases with the median splicing amount of all genes belonging to the supergene in the human kidney. Each triangle represents a supergene. All supergenes have the same total splicing amount.

(D) Spearman's correlation between the median splicing amount of a supergene and its back-splicing rate in each tissue of each mammal examined. All correlations have $p < 0.01$. (E) The back-splicing rate in the human kidney of the paralog with the relatively low splicing amount tends to exceed that of the paralog with the relatively high splicing amount within a paralogous gene pair. The original data are used here. Each dot represents a paralogous gene pair. Dots above and below the diagonal are colored red and blue, respectively. Numbers of red and blue dots are indicated with the corresponding color. The p value is from a binomial test of the null hypothesis of equal numbers of red and blue dots. (F) Proportion of paralogous gene pairs for which the back-splicing rate of the paralog with the relatively low splicing amount exceeds that of the paralog with the relatively high splicing amount in each tissue of each species examined. Both original (squares) and downsampled (triangles) data are used. All fractions are significantly greater than the chance expectation of 50% ($p < 10^{-4}$).

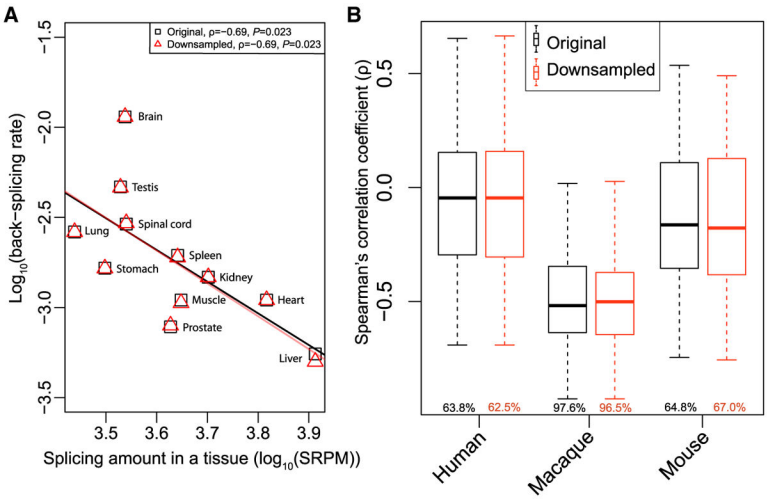


Figure 3. Negative correlation between the back-splicing rate and splicing amount across tissues

The back-splicing rate is the number of back-spliced reads divided by the number of spliced reads in a supergene.

(A) The back-splicing rate of a particular supergene in a tissue decreases with the total splicing amount of the genes belonging to the supergene in the tissue. The virtually superimposed black and red lines are the linear least-squares regressions for the original and down-sampled data, respectively.

(B) Distribution of Spearman's correlation coefficient between back-splicing rate and splicing amount across tissues for all supergenes. In each boxplot, the lower and upper edges of a box represent qu1 and qu3 quartiles, respectively; the horizontal line inside the box indicates the md; and the whiskers extend to the most extreme values inside inner fences, md ± 1.5(qu3 − qu1). Below each boxplot is the fraction of supergenes showing a negative correlation; all fractions significantly exceed 50% ($p < 0.05$, binomial test).

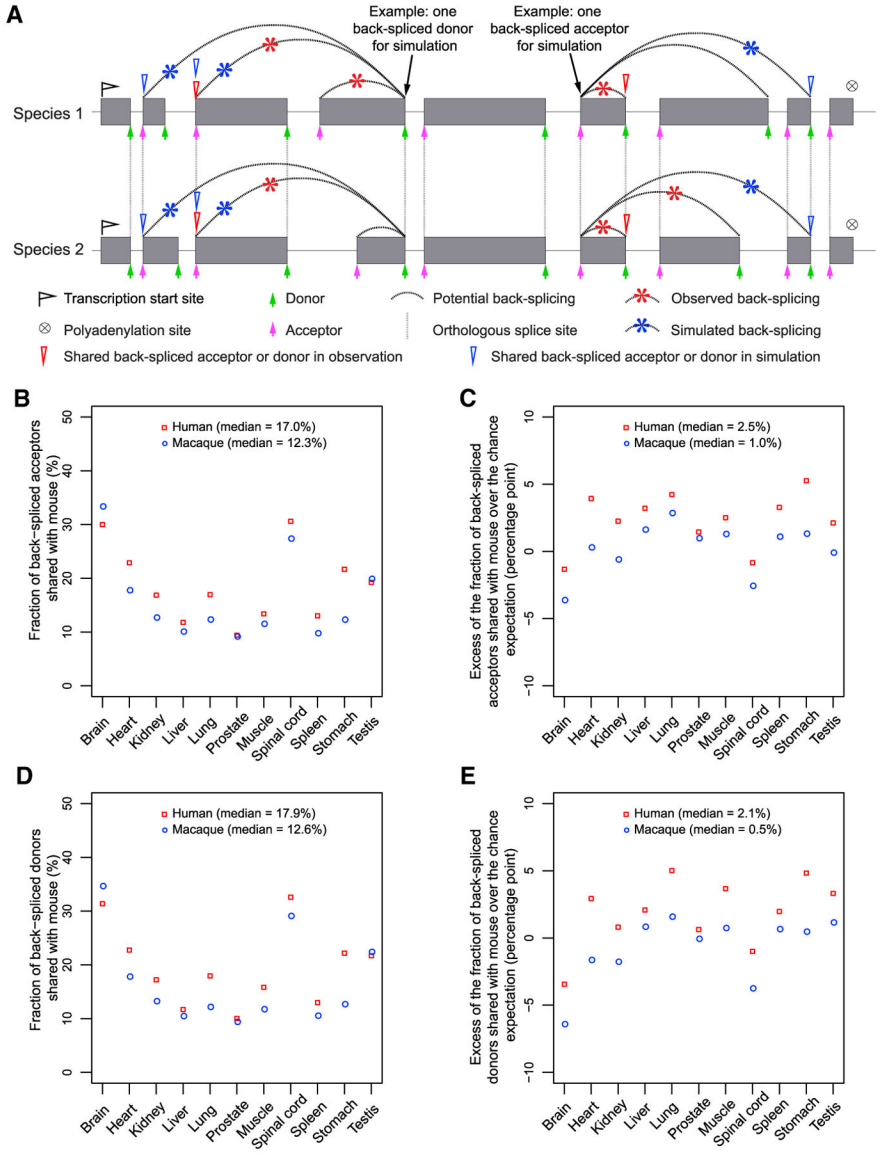


Figure 4. Fractions of human or macaque back-spliced acceptors or donors that are shared with mouse, see also Data S1 and S2

(A) A diagram illustrating various terms used in the analysis. Dotted curves show all potential back-splicing, while red and blue stars indicate realized back-splicing in observation and simulation, respectively.

(B) Fraction of human or macaque back-spliced acceptors shared with mouse.

(C) Difference between the fraction of back-spliced acceptors shared with mouse and the chance expectation.

(D) Fraction of human or macaque back-spliced donors shared with mouse.

(E) Difference between the fraction of back-spliced donors shared with mouse and the chance expectation.

In (B)–(E), the median value across the 11 tissues is provided in the parentheses after each species.

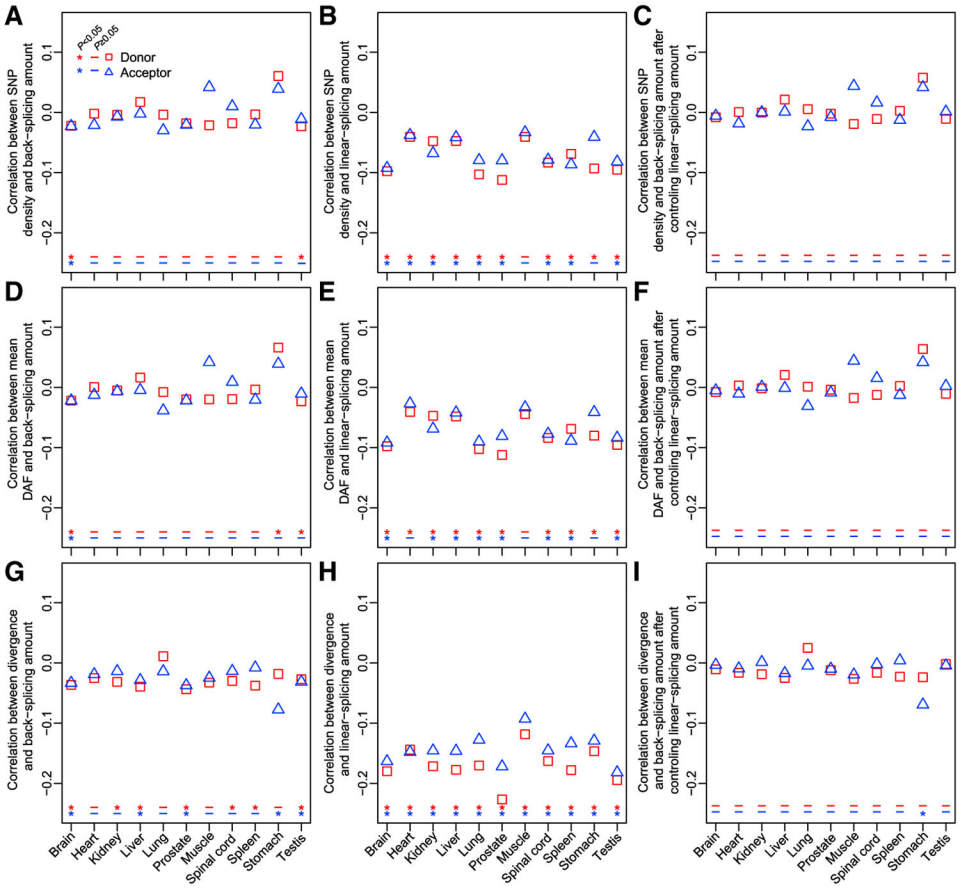


Figure 5. Back-splicing signals are not protected by purifying selection

(A) Spearman's correlation between human SNP density at a splice-acceptor or donor site and the associated back-splicing amount.

(B) Spearman's correlation between human SNP density at a splice-acceptor or donor site and the associated linear-splicing amount.

(C) Partial rank correlation between human SNP density at a splice-acceptor or donor site and the associated back-splicing amount, upon the control of the linear-splicing amount.

(D–F) Same as (A)–(C) except that human SNP density is replaced with human mean derived allele frequency (DAF).

(G–I) Same as (A)–(C) except that human SNP density is replaced with human-macaque divergence. Statistical significance of a correlation is indicated by a dash (non-significant) or star (significant at p = 0.05) at the bottom of each panel.

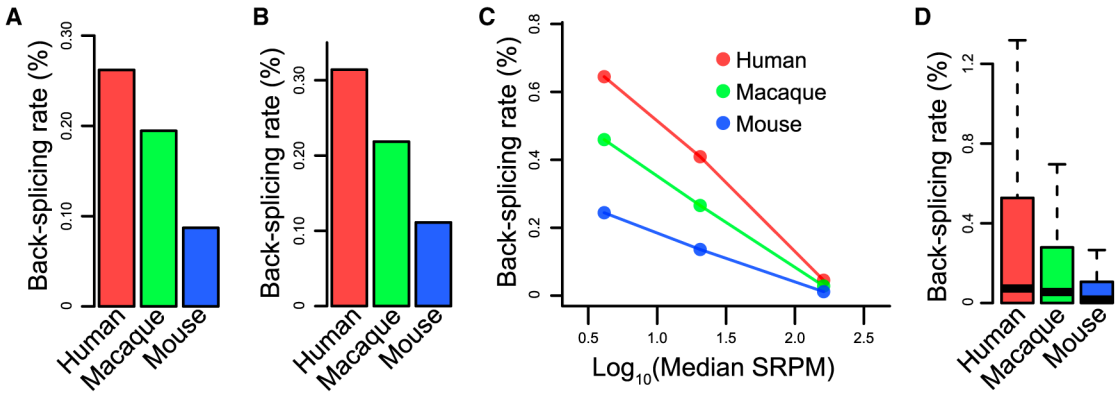


Figure 6. The overall back-splicing rate in a species declines with its effective population size

(A) Overall back-splicing rates in the human, macaque, and mouse when all (expressed and spliced) genes are considered.

(B) Overall back-splicing rates in the three species when only one-to-one orthologous genes are considered. In (A) and (B), the difference between any two species is significant ($p < 10^{-15}$, Fisher's exact test).

(C) Among-species comparison of back-splicing rates of one-to-one orthologous genes after the genes are stratified into bins of low (<10 SRPM), intermediate (10 to 100 SRPM), and high (> 100 SRPM) splicing amounts according to the mean splicing amount across species. SRPM, number of spliced reads per million total reads in a sample. At each of the three levels of splicing amount, the difference in back-splicing rate between any two species is significant ($p < 10^{-15}$, Fisher's exact test).

(D) Boxplot showing the distribution of the back-splicing rate among one-to-one orthologous genes in each species. The difference between any two species is significant ($p < 10^{-238}$, Wilcoxon signed-rank test). In each boxplot, the lower and upper edges of a box represent qu1 and qu3, respectively; the horizontal line inside the box indicates the md; and the whiskers extend to the most extreme values inside inner fences, md ± 1.5(qu3 − qu1).

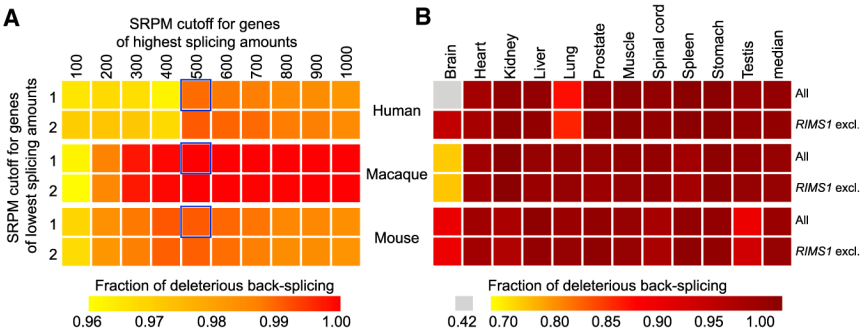


Figure 7. Fraction of deleterious back-splicing

(A) Estimated fractions of deleterious back-splicing for each species when data from all tissues are combined, under different cutoffs for genes of the lowest and highest splicing amounts.

(B) Estimated fractions of deleterious back-splicing in each tissue of each species examined under the cutoffs boxed in (A), either when all genes are considered or when *RIMS1* is excluded. Back-splicing of *RIMS1* in the human brain is unusually abundant and may be beneficial.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RiboMinus RNA-seq data	Ji et al., 2019	https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA000751
RNA-seq data	Ji et al., 2019	https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA000751
RNase R+ data	Ji et al., 2019	https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA000751
Polymorphism data	IGSR	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/
Genome assembly and annotation data	ENSEMBL	http://may2017.archive.ensembl.org/index.html
Software and algorithms		
R	The R Foundation	https://www.r-project.org/
Perl	The Perl Foundation	https://www.perl.org/
Python	Python Software Foundation	https://www.python.org/
LiftOver	UCSC	https://genome.ucsc.edu/util.html
BioMart	ENSEMBL	http://useast.ensembl.org/biomart/martview//38de6b77bceb11d76acd2a1d1b231382
BWA	Li and Durbin 2009	http://bio-bwa.sourceforge.net/
Tophat2	Kim et al., 2013	https://ccb.jhu.edu/software/tophat/index.shtml
STAR	Dobin et al., 2013	https://github.com/alexdobin/STAR
CircExplorer2	Zhang et al., 2016	https://circexplorer2.readthedocs.io/en/latest/
CIRI2	Gao et al., 2018	http://159.226.67.237:8080/new/download_file.php