Open
LETTER TO THE EDITOR

Cell Research (2016) :1-4.
www.nature.com/cr

CircRNA-derived pseudogenes

Cell Research advance online publication 29 March 2016; doi:10.1038/cr.2016.42

Dear Editor,

Circular RNAs (circRNAs) from pre-mRNA back-splicing have been recently re-discovered genome-wide in metazoans by taking advantage of RNA deep-sequencing (RNA-seq) of the non-polyadenylated transcriptomes and specific computational pipelines that can retrieve the non-colinear back-splicing junction reads [1-5]. Despite being inefficiently processed and mostly expressed at a low level [1, 2, 6-8], some circRNAs are derived from genomic loci associated with human diseases [9], and increasing lines of evidence have suggested their potential roles in transcriptional, post-transcriptional and translational regulation [10]. However, nothing is known about whether these exceptionally stable circRNAs can be retrotranscribed, and ultimately inserted back into the host genome as processed pseudogenes.

Theoretically, a linear mRNA-derived pseudogene keeps the same sequential order (colinear) of exons as in its parent linear mRNA (Figure 1A). In contrast, a circRNA-derived pseudogene would have an exon-exon junction in a reversed order (non-colinear) (Figure 1B). By taking advantage of this feature, we developed a computational pipeline (CIRCpseudo) to identify potential circRNA-derived pseudogenes in the mouse reference genome (Figure 1C and Supplementary information, Data S1). Among them, at least 33 pseudogenes are possibly derived from the same circular RNA at the *RFWD2* (ring finger and WD repeat domain 2) locus (*circRFWD2*) with the characteristic non-colinear back-splicing junction sequences that anchor exon 6-exon 2 in a reversed order (Figure 1D and Supplementary information, Table S1A). We referred to these 33 pseudogenes as "high-confidence *circRFWD2*-derived pseudogenes" owing to the existence of the non-colinear exon 6-exon 2 junction sequence.

The number of circRNA-derived pseudogenes identified by the CIRCpseudo pipeline could be underestimated due to the lack of back-splicing junction sequences in the pseudogene loci. For example, although a circRNA-derived pseudogene could be produced from a circRNA, if its back-splicing junction sequence was either not reverse transcribed to cDNA prior to its insertion into the genome, or reverse transcribed to cDNAs and inserted into the genome but lost during evolution, it cannot be revealed by CIRCpseudo. Supporting this notion, when mapping linear *RFWD2* cDNA sequence to the mouse reference genome, another nine *RFWD2*-related pseudogenes were also found in the mouse genome. Different from the high-confidence *circRFWD2*-derived pseudogenes with the non-colinear exon 6-exon 2 junction sequences, these nine pseudogenes contain sequences only in *circRFWD2*-residing exons, i.e., exons 2 to 4 or 5, but lack the back-splicing exon 6-exon 2 junction sequences and the sequences outside of *circRFWD2*. Thus, we referred this subgroup of nine pseudogenes as "low-confidence *circRFWD2*-derived pseudogenes" (Supplementary information, Table S1A). In addition, we detected six more *RFWD2*-related pseudogenes with sequences beyond *circRFWD2*-residing exons (Supplementary information, Table S1A), which suggests that these six *RFWD2*-related pseudogenes might have originated from retrotransposition of (linear) *RFWD2* mRNA.

Although most, except two high-confidence and one low-confidence, *circRFWD2*-derived pseudogenes (Supplementary information, Table S1A) have been annotated previously, none was annotated as retrotransposed from the *RFWD2* circRNA. The finding of *circRFWD2*-derived pseudogenes thus suggests a previously unknown impact of circRNAs: to insert into the genome and change genomic DNA composition. Interestingly, these *circRFWD2*-derived pseudogenes contain different lengths of *circRFWD2*-related sequences with different mutation rates (Figure 1D and data not shown), suggesting different insertion times and/or distinct evolutionary constraints imposed on these *circRFWD2*-derived pseudogenes. All 33 high-confidence *circRFWD2*-derived pseudogenes could be found in the genomes of eight other mouse strains examined but not in rat and primate genomes (Supplementary information, Figure S1A and data not shown). This suggests that the retrotransposition of *circRFWD2* occurred in the common mouse ancestor after its divergence from rat and human lineages.

It is generally believed that pseudogenes processed from (linear) mRNAs contain a canonical structural feature of having a poly(A) tail at the 3′ end [11], and a 3′

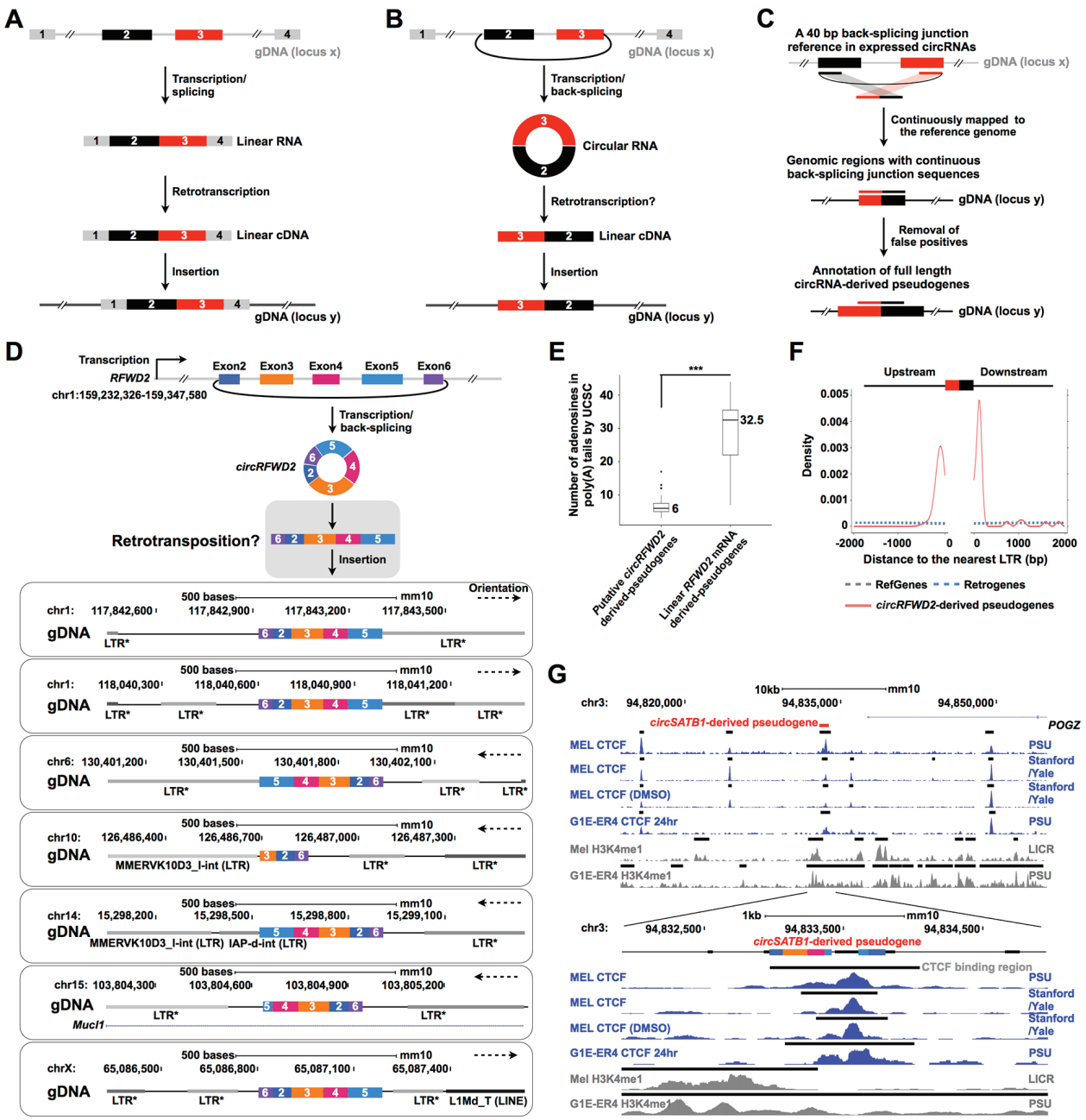

Figure 1 Characterization of circRNA-derived pseudogenes and their potential role in reshaping genome architecture. **(A)** A schematic diagram of the generation of a pseudogene from a (linear) mRNA. **(B)** A schematic diagram of the generation of a circRNA-derived pseudogene. **(C)** Genome-wide identification of circRNA-derived pseudogenes by CIRCpseudo. A reference containing 40 bp back-splicing junction sequences (20 bp on either side of junction) in expressed circRNAs was constructed, and then mapped to the genome to identify circRNA-derived pseudogenes (Supplementary information, Data S1). **(D)** Characterization of mouse *circRFWD2*-derived pseudogenes. A *circRFWD2* that contains exons 2, 3, 4, 5 and 6 with back-splicing of the exon 6-exon 2 junction sequence was produced from the mouse chr1:159 232 326-159 347 580 locus (top), and could be retrotransposed (middle, gray box) to generate pseudogenes at different genomic regions (bottom). *, MMERVK10C-int LTR retrotransposon sequences. **(E)** Counts of adenosines in 3′ ends of poly(A) tails in all (both linear and circular) *RFWD2*-originated pseudogenes by UCSC RetroGene annotation. 39 out of 42 putative *circRFWD2*-derived pseudogenes annotated by UCSC have significantly fewer adenosines than those in the six linear *RFWD2* mRNA-derived pseudogenes. *** $P = 6.0 \times 10^{-4}$, Wilcoxon rank-sum test. **(F)** LTRs are highly enriched in the flanking regions of the *circRFWD2*-derived pseudogenes. The nearest LTRs are significantly higher in the flanking regions of all 42 *circRFWD2*-derived pseudogenes (red solid line) than those in mouse RefGenes (gray dashed line) or mouse RetroGenes (blue dashed line). **(G)** A CTCF-binding site resides

poly(A) sequence was recently reported to be required for LINE-1-mediated retrotransposition in human [12]. We therefore examined the adenosine numbers within UCSC-annotated poly(A) tails of all pseudogenes that have mouse *RFWD2* origin (Supplementary information, Table S1A). Interestingly, 39 out of the 42 *circRFWD2*-derived pseudogenes (annotated as pseudogenes also by UCSC) contain only a few adenosines in their UCSC-annotated 3′ end poly(A) tails in general; in contrast, the six linear *RFWD2* mRNA-derived pseudogenes on average are enriched with adenosines in their 3′ ends (Figure 1E and Supplementary information, Table S1A). This further suggests that *circRFWD2*-derived pseudogenes might have originated from circRNAs that in general rarely have canonical poly(A) tails. In addition, mouse LTR sequences are highly enriched in the flanking regions of 42 mouse *circRFWD2*-derived pseudogenes, compared to those that flank the annotated RefGenes and RetroGenes (Figure 1F). This analysis indicates that the retrotransposition of the *circRFWD2*-derived pseudogenes might be associated with retrotransposons in mouse. The detailed mechanism of circRNA retrotransposition remains mysterious.

Another two high-confidence and dozens of low-confidence circRNA-derived pseudogenes were additionally identified in the mouse reference genome (Supplementary information, Figure S1B and Table S1B). Interestingly, the mouse *circSATB1*-derived pseudogene could be found in all examined mouse strains and the rat reference genome (Supplementary information, Figure S1B, left panel), suggesting that the retrotransposition of *circSATB1*-derived pseudogene might have occurred in the common ancestor of mice and rats. In contrast, the mouse *circDIAP3*-derived pseudogene homolog could be identified in some but not all examined mouse strains (Supplementary information, Figure S1B, right panel), suggesting that the retrotransposition of *circDIAP3*-derived pseudogene might have occurred after the divergence of different mouse strains.

We also identified a few cases of high-confidence and dozens of low-confidence circRNA-derived pseudogenes using a similar strategy in human genomes (Supplementary information, Figure S1C and Table S1C). Interestingly, the homologous sequences of human *circPRKDC*-derived and *circCAMSAP1*-derived pseudogenes could be found in the gorilla and chimp genomes but not in the rhesus genome (Supplementary information, Figure S1C), indicating that their retrotransposition might have happened very recently in evolution. It is worth noting that some pseudogenes, although having non-colinear exon-exon junctions, could be produced by other mechanisms, such as reverse transcription that could go through open-ended but highly-structured RNAs with close enough 5′ and 3′ ends (Supplementary information, Figure S1D).

Although the total number is low so far, the finding of circRNA-derived pseudogenes suggests a previously underestimated impact of circRNAs on host genome by retrotransposition. Are these insertions of retrotransposed circRNAs functional? Expressed pseudogenes have been suggested to play important roles in cellular differentiation and cancer progression [13]. The detection of expressed circRNA-derived pseudogenes could be limited by the sequence similarity between circRNA-derived pseudogenes and their original circRNAs, especially at the non-colinear back-splicing junction regions, which are key to the annotation of circRNAs. For example, RNA-seq reads that are mapped to *circRFWD2* are similar to the sequences from *circRFWD2*-derived pseudogenes (Supplementary information, Figure S1E).

In addition to their potential functions at the RNA level after transcription, we speculated that the insertion of retrotransposed circRNAs might be involved in gene expression regulation by altering the genome structure. Intriguingly, a CTCF/Rad21-binding site in the mouse MEL and G1E cell lines was identified to overlap exactly with *circSATB1*-derived pseudogene locus (Figure 1G and data not shown). This area has also been suggested as an enhancer region with active H3K4me1 signals. The CTCF binding is specific for the *circSATB1*-derived pseudogene region, but not its original *SATB1* region of exons 2, 3, 4 and 5 for *circSATB1* (Supplementary information, Figure S1F). As CTCF binding could affect chromosome configuration and thus regulate gene expression [14], this finding may indicate an unexpected biological effect originated from circRNAs. Strikingly, the *circSATB1*-derived pseudogene could not be found in the examined human genomes, neither could the corresponding CTCF-binding site be found in available CTCF datasets (Supplementary information, Figure S1G).

This study demonstrates that pseudogenes can be retrotransposed from circRNAs and, importantly, inherited

in the mouse *circSATB1*-derived pseudogene region in the mouse ENCODE MEL and G1E cell lines. Correspondingly, this area is also suggested as an enhancer region with active H3K4me1 signals. Blue peaks, CTCF-binding signals; gray peaks, H3K4me1-binding signals; black bars over the binding signals, predicted CTCF/H3K4me1-binding regions.

in mammalian genomes. Their existence in the genome has the potential to reshape genome architecture by providing additional CTCF-binding sites. Further efforts are needed to decipher the molecular mechanism of circRNA retrotransposition, annotate circRNA-derived pseudogenes in different species, profile their expression patterns in various transcriptomes, and demonstrate what other unexpected roles they could play in cell.

Acknowledgments

We are grateful to Gordon Carmichael for reading the manuscript. Next generation deep sequencing was performed at the CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China. This work is supported by grants 2014CB964800 and 2014CB910600 from the Ministry of Science and Technology of China, and grants 91540115, 91440202 and 31471241 from the National Natural Science Foundation of China.

Rui Dong[1], Xiao-Ou Zhang[1], Yang Zhang[2], Xu-Kai Ma[1], Ling-Ling Chen[2, 3], Li Yang[1, 3]

[1]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; [2]State Key Laboratory of Molecular Biology, Institute of Biochemistry and Cell Biology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; [3]School of Life Science, ShanghaiTech University, Shanghai 200031, China
Correspondence: Li Yang
E-mail: liyang@picb.ac.cn

References

1. Jeck WR, Sorrentino JA, Wang K, *et al*. *RNA* 2013; **19**:141-157.
2. Salzman J, Chen RE, Olsen MN, *et al*. *PLoS Genet* 2013; **9**:e1003777.
3. Westholm JO, Miura P, Olson S, *et al*. *Cell Rep* 2014; **9**:1966-1980.
4. Zhang XO, Wang HB, Zhang Y, *et al*. *Cell* 2014; **159**:134-147.
5. Ivanov A, Memczak S, Wyler E, *et al*. *Cell Rep* 2015; **10**:170-177.
6. Guo JU, Agarwal V, Guo H, *et al*. *Genome Biol* 2014; **15**:409.
7. Starke S, Jost I, Rossbach O, *et al*. *Cell Rep* 2015; **10**:103-111.
8. Zhang Y, Wei X, Li X, *et al*. *Cell Rep* 2016; In press.
9. Burd CE, Jeck WR, Liu Y, *et al*. *PLoS Genet* 2010; **6**:e1001233.
10. Chen LL. *Nat Rev Mol Cell Biol* 2016 Feb 24; doi:10.1038/nrm.2015.32
11. Vanin EF. *Annu Rev Genet* 1985; **19**:253-272.
12. Doucet AJ, Wilusz JE, Miyoshi T, *et al*. *Mol Cell* 2015; **60**:728-741.
13. Kalyana-Sundara S, Kumar-Sinha C, Shankar S, *et al*. *Cell* 2012; **149**:1622-1634.
14. Tang Z, Luo OJ, Li X *et al*. *Cell* 2015; **163**:1611-1627.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)