Supplementary Methods

Identification of all possible back-splicing junctions from RNA-seq datasets

CIRCexplorer [*1*] was first used to annotate circRNAs from 19 mouse cell lines (GEO:GSE65926) and mouse embryonic stem cell R1 with/without RNase R poly(A)– (GEO:GSE53942 and GSE60467). Another method was additionally set up to annotate circRNAs. Briefly, a back-splicing junction reference was constructed by allowing 95 bp of annotated (Mouse: knownGene.txt updated at 2015/06/01, refFlat.txt updated at 2015/07/29 and ensGene.txt updated at 2014/04/06) exon sequences on either side of back-splicing junction. Then, 100 bp (or trimmed to 100 bp) RNA-seq reads from the same mouse datasets were mapped to this back-splicing junction reference with at least 5 bp alignment overhang from one side of junction by using Bowtie (v0.12.9, -m 1 -v 1). Finally, the back-splicing junctions within all expressed mouse circRNAs identified by both CIRCexplorer method and the new method described above were obtained.

Human circRNAs were identified with similar strategy from 15 ENCODE cell line (GEO:GSE26284), six neuron differentiation samples (GEO:GSE65926), and RNase R treated or untreated PA1 and H9 poly(A)–/ribo– RNA-seq (GEO:GSE75733, GSE24399, GSE48003, GSE60467).

A computational pipeline for circRNA-derived pseudogenes

A stepwise pipeline was developed to comprehensively annotate circRNA-derived pseudogenes by identifying the non-colinear back-splicing junction sequences continuously in the examined reference genomes (CIRCpseudo). Briefly, a 40 bp back-splicing junction reference from all identified circRNAs was constructed by spanning at

least 20 bp sequences on each side of back-splicing junction. After removing false positives that can overlap with colinear exon-exon junctions and exon-intron junctions, the 40 bp back-splicing junction sequences were mapped by BWA (v0.6.2, -n 4) to the mouse reference genome continuously at the genomic DNA level to find all possible genomic loci with non-colinear 40 bp back-splicing junction sequences. Based on these identified 40 bp back-splicing junction sequences, full length of back-spliced exon sequences were fetched form annotated circRNAs and further mapped to the reference genome by BLAT with the default setting to identify circRNA-derived pseudogenes and remove false positives. All the identified circRNA-derived pseudogenes were then fully annotated by comparing with their original circRNAs to profile start positions, end positions, the sequence compositions and exon boundaries.

To find circRNA-derived pseudogenes with shorter back-splicing junction sequences on one side, 30 bp mouse colinear exon-exon junctions (15 bp on each exon) were alternatively constructed and applied to similar analysis, leading to the identification of circRNA-derived pseudogenes with shorter junction sequences.

The source codes of CIRCpseudo pipeline and related references/documents can be accessed from https://github.com/YangLab/CIRCpseudo.

Classification of different types of pseudogenes

The high-confidence circRNA-derived pseudogenes have the non-colinear back-splicing junction sequences in their genomic DNA level, and in contrast, the low-confidence circRNA-derived pseudogenes contain only circRNA-residing exon sequences, but lack the back-splicing junction sequences in the genome.

Manual annotation of additional circRNA-derived pseudogenes

To discover the low-confidence circRNA-derived pseudogene candidates without fusion junctions, sequences of the identified circRNAs were retrieved and then directly aligned to the mouse references genome by BLAT with default setting. Mapped sequences with ≥200 bp in length and ≤500 bp gap were kept for further checking. Extra circRNA-derived pseudogenes without back-splicing junction sequences were further identified in the examined mouse reference genome.

Poly(A) and LTR distribution analysis of pseudogenes

Counts of As in annotated poly(A)s of all *RFWD2* (both *circRFWD2* and linear *RFWD2* mRNA) derived pseudogenes were retrieved from RetroGene annotation of UCSC (v6, ucscRetroInfo6.txt updated at 2015/03/02). And LTR distribution was obtained from ReapeatMasker file (rmsk.txt updated at 2012/03/07).

Other used genomes from different mouse strains and human populations

Genomes of eight mouse strains (129P2/OlaHsd, 129S5SvEvBrd, AKR_J, BALB_cJ, BuB_BnJ, C3H_HeH, C57BL_6NJ and C57BR_cdJ) were retrieved from Mouse Genome Project of Sanger Institute (http://www.sanger.ac.uk/science/data/mouse-genomes-project) and upload to UCSC for visualization.

Ten human populations (ACB: African Caribbean in Barbados, ASW: African Ancestry in Southwest US, BEB: Bengali in Bangladesh, CDX: Chinese Dai in

Xishuangbanna, China, CEU: Utah residents with Northern and Western European ancestry, CHB: Han Chinese in Bejing, China CHS: Southern Han Chinese, China CLM: Colombian in Medellin, Colombia ESN: Esan in Nigeria and FIN: Finnish in Finland) from 1000 Genomes Project (http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/) [*2*] were selected for circRNA-derived pseudogenes analysis in human.

CTCF binding analysis

ChIP-seq datasets of CTCF and H3K4me1 in mouse Mel, G1E-ER4 cell lines (GEO: GSE36028, GSE36029, GSE31039; UCSC accession: wgEncodeEM001947, wgEncodeEM001968, wgEncodeEM001983, wgEncodeEM002781) were mapped to the mouse mm10 reference genome by Bowtie (0.12.9) with up to 2 mismatches (parameters: -m 1 -v 2). ChIP regions were called by MACS (v1.4.2) [*3*] with bandwidth=125. CTCF binding signals and peaks in related human cell lines were obtained from ENCODE project [*4*]. Conservation of genome coordinates around *circSATB1*-derived pseudogene regions between mouse genome (mm10) and human genome (hg19) was based on liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) with default setting.

References

1. Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, Yang L. *Cell* 2014; **159**:134-147.
2. Clarke L, Zheng-Bradley X, Smith R, *et al. Nat Methods* 2012; **9**:459-462.
3. Zhang Y, Liu T, Meyer CA, *et al. Genome Biol* 2008; **9**:R137.
4. Rosenbloom KR, Sloan CA, Malladi VS, *et al. Nucleic Acids Res* 2013; **41**:D56-63.