Exploration of transposon-derived chimeric transcripts expressed in mammalian embryos.

Saki Kawakami, , Shuntaro Ikeda, Shinnosuke Honda

Reproductive Biology, Graduate School of Kyoto University, Japan

[Introduction]

Transposons are mobile DNA elements and constitute ~40% of mammalian genome. During zygotic genome activation (ZGA), some of transposons are activated and form chimeric transcripts with downstream genes in mammalian embryos though their activation is suppressed generally. Although some of these chimeric mRNAs have been found to have different functions from the canonical transcripts and contribute to cell proliferation and differentiation in mouse embryos, high-quality chimeric mRNA libraries have not yet been constructed and many of them may remain undiscovered. This could be due to the technical limitations of short-read sequencing. Although short-read sequencing is mainstream sequencing method due to its high accuracy, mapping transposon-derived sequences is difficult because they exist in numerous numbers in the genome. Here, we applied long-read RNA-sequencing (RNA-seq) on mouse and bovine ZGA stage embryos to comprehensively search for chimeric transcripts involved in embryonic development. In addition, we performed comparative analysis of short- and long-read sequence data to verify differences in detection and quantification of transcripts.

[Material and Methods]

Two hundred mouse or bovine embryos at the ZGA stage were collected and mRNA was extracted. The mRNA was reverse transcribed to cDNA and sequenced based on the cDNA-PCR Sequencing Kit from Oxford Nanopore Technology Inc. After base calling and quality checking, reads were mapped to the mm10 or bosTau9 reference genome. The short-read RNA-seq data of mouse and bovine embryos were downloaded from the GEO dataset (GSM1845301and GSM4275382) for comparative analysis. After mapping to RefSeq transcript datasets and quantification, Pearson correlation analysis was performed. Expression levels were normalized by RPT10K (Reads Per Transcripts per 10 K mapped

reads). To compare the number of detected transcripts between short- and long-read sequences, we extracted those with RPT10K more than 0 from the quantified data.

[Results]

In both mouse and bovine, several chimeric transcripts listed in previous studies were detected. Especially, Prmt6 and Cdk2ap1 were identified as chimeric transcripts in common with mouse and bovine. We recaptured the phenomenon that the type of transposon used as promotor for chimeric transcripts differed between species; for example, Cdk2ap1 was transcribed from the LTR elements of endogenous retrovirus in mouse, while in bovine it was transcribed from the fusion region of the DNA transposon and LINE.

Comparative analysis between short- and long-read sequencing did not show high correlation coefficients. These results could be due to incorrect mapping of transposon-derived reads in short-read sequencing, or due to overestimation of quantitative values caused by PCR or normalization bias in long-read sequencing.

In addition, there was no significant difference in the number of transcripts detected between short- and long-read sequences in both mouse and bovine. Especially, comparison of bovine sequencing data showed that more genes were detected in the long reads than in the short reads. This suggests that long-read sequencing is suitable for searching for novel chimeric transcripts.

Our research suggests the possibility that long-read sequence makes it possible to discover novel chimeric transcripts more comprehensively.