A landscape of full-length RNAs in human

Shruti Bhagat[1*], Shoya Kato[1,2*], Zhiwei Zhang[1,3*], Kazuhiro Takeuchi1[2,4*], Sho Sekito[1,5], Fumiya Wada[1,6], Akiko Oguchi[1,4,7], Hideya Kawaji[8], Yasuhiro Murakawa[1,2,4,9]

[1] *Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University.*
[2] *Department of Medical Systems Genomics, Graduate School of Medicine, Kyoto University.*
[3] *Department of Human Genetics, Faculty of Medicine and Health Science, McGill University.*
[4] *RIKEN-IFOM Joint Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences.*
[5] *Department of Nephro-Urologic Surgery and Andrology, Graduate School of Medicine, Mie University.*
[6] *Department of Hematology, Graduate School of Medicine, Kyoto University.*
[7] *Department of Nephrology, Graduate School of Medicine, Kyoto University.*
[8] *Research Center for Genome & Medical Sciences, Tokyo Metropolitan Institute of Medical Science.*
[9] *IFOM ETS - the AIRC Institute of Molecular Oncology.*
** These authors contributed equally.*

Abstract

Identification of isoforms and genes in humans remains incomplete owing to the limitations of conventional technologies. Here, we present a new long-read RNA-seq method (FLAM-seq2), which elucidates the complete structure of RNAs from 5' to 3' end with single-molecule sequencing. We studied full-length RNAs in 43 normal human tissues and cancer cell lines in both total and cytoplasmic RNA fractions amounting to ~0.5 billion sequenced reads. The median length of our reads is 3,000 bp, which captures a large portion of mRNAs. Using our original computational framework, we identified over one million alternate isoforms for known genes with different splice junctions and transcription start and end sites. These alternate isoforms can alter known open reading frames. Additionally, we identified close to 20,000 novel candidate genes, many of which are detected in the cytoplasmic fraction of cancer cell lines. Integration with deep proteomics data supports the presence of potential protein-coding genes. Interestingly, many novel genes are conserved in primates but not in mammals. In summary, we provide a broad and deep resource of full-length RNAs that have widespread applications in biomedical research and therapeutics.