Pan-cancer analysis reveals roles of retrotransposon-fusion RNAs

Junseok Park

Transposons make up about half of the human genome, and some full-length elements create novel insertions in germ cells and somatic tissue of both healthy and diseased individuals. These genetic elements have the potential to generate chimeric transcripts consisting of transposons and non-transposons, which may result in the production of abnormal proteins or immunogenic molecules. However, accurate identification of these transposon fusion events from short RNA-seq reads is challenging due to the repetitive nature of transposon sequences and read alignment errors near exon-intron junctions. To address this, we have developed a computational pipeline, rTea (RNA Transposable Element Analyzer) to detect various types of transposons-fusion transcripts from reference and non-reference, i.e., polymorphic or somatic transposon insertions from RNA-seq data. We applied rTea to analyze 10,257 cancer samples across 34 cancer types, as well as 3,088 normal tissue samples from the TCGA/ICGC, our unpublished colorectal cancer cohort, and the GTEx consortia. We realigned all RNA-seq reads for unified processing on Google Cloud Platform (GCP) using GenomeFlow, a tool to design scalable distributed processing architectures to optimize computational resources and reduce compute costs. We identified 30,016 fusions with an average of 203 events per normal sample, particularly abundant in the testis. We also identified 52,277 cancer-specific fusions that were not detected in the corresponding normal tissues, with an average of 30 events per cancer sample. We found that the somatic cancer fusions were enriched in known cancer genes, suggesting their involvement in tumorigenesis. Furthermore, we discovered distinct splicing hotspots and DNA methylation changes associated with fusions from different families of source transposons. Our in silico immunogenicity modeling and experimental validation confirmed that several peptides derived from transposon fusions in cancers bind to MHC-I and activate CD8+ T cells to a comparable extent to EBV viruses. Our findings highlight the potential of endogenous retroelements as novel therapeutic targets and a significant source of neoantigens. rTea is available at https://github.com/ealeelab/rtea.