Mapping functional non-coding transcripts in primate immune cells using comparative epigenomics

David Venuto[1], Jan Paul Buschdorf[2], Shu-Huang Chen[3], Maxime Caron[4], Tony Kwan[3], Michael Meaney[2,5], Tomi Pastinen[1,3], Guillaume Bourque[1,3]

[1] *Department of Human Genetics, McGill University, Montréal, Canada*
[2] *Singapore Institute for Clinical Sciences, Singapore, Singapore*
[3] *McGill University and Genome Québec Innovation Centre, Montréal, Canada*
[4] *Centre de Recherche du CHU Sainte-Justine, Montréal, Canada*
[5] *Douglas Mental Health Institute, McGill University, Montréal, Canada*

Abstract

It is known that although only 2% of the human genome encodes for proteins, a much larger portion is transcribed. Transposable elements (TEs), which comprise about 45% of the human genome, have been shown to contribute a significant portion of these expressed non-coding transcripts. While some TE-derived transcripts have been functionally annotated, the majority have only been studied in a limited number of cell types and are largely uncharacterized in terms of genomic function. To help identity functional non-coding transcripts in human immune cells, we generated matched RNA-seq datasets in human (*Homo sapiens*) Monocytes, T-Cells and B-Cells and in macaque (*Macaca mulatta*) Monocytes and T-Cells; and looked for non-coding transcripts that are well conserved in these two primate species. As expected, global gene expression analysis revealed a clustering by cell types and not by species. In contrast, global repeat expression analysis revealed a clustering by species. Differential expression (DE) analysis utilizing summed count values of all instances per repeat family was used to identify repeats with unique expression profiles in the pairwise comparisons. Using this approach, 320 (24.1%) and 350 (25.8%) repeat families were identified as DE between T-cells and Monocytes in human and macaque respectively. In particular, LIM3 type families, LTR52-int, MamGypLTR3, MLT1-int and MER92-int, were found to be differentially expressed between the two species. In contrast, a total of 577 (6.28%) repeat families were identified as having conserved expression across all cell types in both species. This included the Tigger, MER45B, LTR75, MamRep, and Charlie TE families. Next, we looked for additional evidence to support the functionality of some of these non-coding transcripts. First, we looked for the presence of RNA secondary structure (SS) motifs and found that 301 repeat families (had at least one instance of a RNA SS motif conserved using a prediction pipeline that relied on a 13-species sequence comparison. In total, 519 TE instances with such a conserved RNA SS motifs were found to be expressed. Similarly, we also identified expressed TE instances that either harbored known transcription factor binding sites or overlapped anthropoid-specific constrained regions, which provided further evidence of their potential functionality. We anticipate that these results will facilitate the discovery of TE-derived non-coding transcripts that contributed to immune cell innovations in the primate lineage.