HERV copy number variation in the human population

Jainy Thomas and Cedric Feschotte

Department of Human Genetics, University of Utah School of Medicine, Salt Lake City.

Increased transcript levels of human endogenous retroviruses (HERV) have been associated with diseases such as amyotrophic lateral sclerosis and multiple sclerosis. Copy number variation (CNV) is one among multiple factors that may alter gene expression levels. Reinfection or retrotransposition, resulting in the generation of new provirus copies, and recombination between two flanking long terminal repeats (LTRs), resulting in deletion of the internal coding sequence and formation of a solo LTR are features that may enhance HERV CNV. In the case of recently active HERV-K (HML2), these mechanisms have played a role in generating CNVs in human populations. In contrast, most of the HERV-W insertions occurred >25 million years ago and virtually nothing is known about CNVs and the possible contribution of this type of polymorphism to disease susceptibility**.** To investigate the presence of HERV-W CNVs in the human population, we used a dataset comprising of 279 whole genome sequences from 142 diverse populations generated by the Simons Genome Diversity Project. We recovered known HERV-K non-reference insertions from the dataset but did not identify any non-reference HERV-W insertions in the dataset suggesting that germline retrotransposition is unlikely to have played a role in the generation of HERV-W CNV. Next we developed a novel computational pipeline to identify CNV attributable to recombination events between LTRs and identified 12 candidates. Deletion dimorphism for one candidate on chromosome 18 (HERV-W18) was validated using PCR on genomic DNA obtained from seven individuals from four South Asian populations. The results were consistent with the predictions made by our pipeline confirming that HERV-W18 exists as two different allelic forms in the human population, one consisting of a solo LTR and the other being a complete provirus. The allele frequencies of both solo LTR and provirus were calculated based on the genotype predictions and were found to be common (~48-52%) in the human population. Together these data provide, to our knowledge, the first evidence for HERV-W CNV in the human population, raising the possibility that this form of variation could contribute to susceptibility to multiple sclerosis and other diseases associated with HERV-W.