Retroduplications come from reverse transcription of mRNAs and their insertion back into the genome. In this study, we performed comprehensive discovery and analysis of retroduplications in a large cohort of 2,535 individuals from 26 human populations in 1000 Genomes Phase 3. We developed an integrated approach to discover novel retroduplications combining both high-coverage exome and low-coverage whole-genome sequencing data, utilizing information from both exon-exon junctions and discordant paired-end reads. We found 503 parent genes having novel retroduplications absent in the reference genome.

Based solely on retroduplication variation, we built phylogenetic trees of human populations; these represent superpopulation structure well and indicate that variable retroduplications are effective population markers. We further identified 43 retroduplication parent genes that differentiate superpopulations. This group contains several interesting insertion events, including SLMO2 retroduplication and insertion into CAV3, which potentially associates with diseases predisposition. We also found retroduplication occurrence to be associated with a variety of genomic features: (1) Insertion sites were correlated with regular nucleosome positioning; and (2) they, predictably, tend to avoid conserved functional regions, such as genes, but, somewhat surprisingly, also avoid introns. (3) Retroduplications tend to be co-inserted with young L1 elements, indicating of recent retrotranspositional activity, and (4) they have a weak tendency to originate from highly expressed parent genes.

Our investigation provides valuable insight into the functional impact and association with genomic elements of retroduplications. We anticipate our discovery approach and analytical methodology to have applications in a more clinical context, where exome sequencing data is abundant and the discovery of retroduplications potentially improves the accuracy of SNP calling from exomes sequencing.