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. We developed an integrated approach to discover novel retroduplications from both high-coverage exome and low-coverage whole-genome sequencing data, utilizing information on both exon-exon junctions and discordant locations of paired-end reads. We found 503 parent genes having novel retroduplications absent in the reference genome. The set reveals the high-resolution landscape of human germline retroduplication polymorphism and gives us the power to perform extensive analysis of retroduplication variation.

Based solely based on retroduplication variation, we successfully constructed phylogenetic trees of human populations, which represent superpopulation structure and indicate that variable retroduplications are effective population markers. We further identified 43 retroduplication parent genes that differentiate superpopulations. This group contains [[is this right??]]several interesting intragenic insertion events, including the SLMO2 retroduplication and insertion into CAV3, which has potential disease association. We found retroduplication insertion sites were associated with regular nucleosome positioning, co-insertion with young L1 elements, and avoidance of functional elements and, somewhat suprisingly, introns. We also found a weak association with parent gene expression level.

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 discovery of retro-duplications can potentially improve the accuracy of SNP calling from exomes.