Structural variation and the assembly of more complete human genomes

Evan E. Eichler
Department of Genome Sciences and Howard Hughes Medical Institute, University of Washington, Seattle, WA

The discovery and resolution of genetic variation is critical to understanding disease and evolution. I will present our most recent work sequencing diverse human and nonhuman primate genomes using both ultra-long and high-fidelity long-read sequencing technologies. We have developed multiplatform methods to fully phase and assemble diploid genomes without parental data. This allows us to detect and sequence resolve most inversions and copy number variants from several bases up to 50 kbp—the vast majority of which are not routinely characterized by short-read sequencing. We are now developing approaches to characterize some of the largest (>50 kbp) and most complex forms of structural variation mapping to segmental duplications, acrocentric regions, and centromeres. Advances in this area have made possible the first telomere-to-telomere assemblies of genomes providing new biological insights into regions typically excluded from human genetic studies. The recovery of megabase pairs of duplicated sequence and structurally variant sequence absent from the reference genome has led to new genetic associations and the identification of new genes missing from the human reference genome, including the discovery of regions both introgressed and under selection in specific human populations. Long-read sequencing technology is revolutionizing our understanding of human genetic variation, and assembly-based variant discovery, we predict, will be the future of genetic and clinical-based research.