PROJECT SUMMARY

Structural variations (SVs), such as deletions, duplications, inversions and translocations, are among the most significant determinants of human genetic diversity to have been discovered, affecting far more bases than single-nucleotide variants in the genome. The TOPMed Program offers an exciting new opportunity to mine genome sequencing and other -omics datasets from a large cohort of individuals for novel SV discovery, analysis and association with diseases of the Heart, Lung and Blood.. However, SVs are inadequately covered by current computational discovery methods, making it likely that a large proportion of variants associated with human disease remain unidentified or poorly characterized. The overarching objectives of this project are to 1) discover SVs at high resolution and large scale, 2) functionally interpret SV origin and phenotypic effects and 3) associate SVs of high functional impact with disease. We bring together a team of pioneers with a proven record of collaboration and innovation in the field of SV discovery, genotyping and large-scale functional genome analysis. We will develop and integrate novel tools for high-resolution SV discovery and use these to comprehensively profile all types of SVs, including complex SVs, in a large subset of the genomes being sequenced (Aim 1). To examine the functional impact of the identified SVs, we will develop cutting-edge methodologies for functional annotation of variants and characterization of associated biological processes through integration of –omics datasets (Aim 2), which will also enable us to prioritize subsets of SVs for association studies proposed in Aim 3. Finally, we will scale up SV detection and analysis through genotyping of all SVs detected in Aim 1 across the 100K samples of the TOPMed Program, which will provide the necessary statistical power for meaningful genotype-phenotype associations for disease-based SV association studies (Aim 3). Our deliverables will be the largest library of validated SVs discovered in humans, together with an unprecedented and broadly applicable platform of cloud-based pipelines for comprehensive, high-resolution and large-scale SV analysis.