SPECIFIC AIMS

We propose to develop and implement for the TOPMed program an integrative computational platform to discover the link between genomic structural variations (SVs) and diseases of the heart, lung and blood. SVs are genetic variations that include deletions, duplications, insertions, inversions and translocations, all of which can have major impacts on genome function and significantly contribute to human phenotypic variation and disease susceptibility. Despite the importance of SVs for human health, the development of computational methods for their discovery and has proven challenging. SVs are structurally diverse, ranging from "simple" events to complex rearrangements, and most currently available algorithms lack the nucleotide-specificity necessary for analyzing the more complex SVs that have the potential significantly alter genome function. Moreover, complex SVs are disproportionately observed in non-coding regions of the genome, making functional interpretation and analysis challenging. There is accordingly a critical need to develop novel computational strategies for 1) mining complex genomic datasets for SV discovery at high resolution and large scale, 2) accurate genotyping and association of specific (and rare) SVs with disease, and 3) functional interpretation of SV origin and functional effects.

We propose to develop methods that will enable comprehensive analysis of SVs with unprecedented depth and resolution. We bring together a team of scientists with a proven record of collaboration and innovation in the field of SV discovery and large-scale functional genomics analysis. With our combined expertise, we will establish a novel platform for discovering, validating and genotyping complex SV events from the thousands of genomes being sequenced by the various projects of the TOPMed program. By applying this platform to data from the TOPMed program and public repositories such as ENCODE, we will provide insight into the functional relevance of SVs and perform trial association studies in a disease-specific context, genotyping a selection of high-impact SVs in the full cohort of individuals sequenced. We will achieve these goals through the following Specific Aims:

Aim 1. Build an integrative pipeline for large-scale discovery of complex structural variation. We will build an integrated, smart and scalable pipeline of SV-calling algorithms, developed by our group and others, to discover all classes of SVs in a select cohort of individuals being sequenced as part of the TOPMed program. Using breakpoint assembly methods, we will perform *in silico* validation of the SV events and will use the assembled contigs to investigate the complexity prevalent at the breakpoints. These studies will deliver the largest reference library of validated SVs discovered in humans and will allow us to make novel biological inferences in the various disease cohorts.

Aim 2. Develop tools to analyze the functional impact of SVs. We will develop a framework to evaluate SVs that impact protein-coding genes, non-coding RNAs and non-coding regulatory regions. This framework will integrate information about evolutionary conservation, existing genomic annotations and epigenetic/transcriptomic datasets from the TOPMed program and other public sources to assign a ***Functional Impact score*** to each SV. In particular, we will up-weight the impact score of SVs that overlap elements with ubiquitous activity, high network connectivity (i.e., hubs), strong allelic activity and eQTL associated variants (i.e., functional sensitivity to variants). Our score will also take into the varied ways that a SV can interact with a functional element (i.e., engulf vs. partially overlap).

Aim 3. Association of structural variants with common and rare diseases. We will genotype the high impact SVs across a larger cohort and we will perform disease phenotype association analysis for high impact SVs. We anticipate that many high-impact SVs will be rare, necessitating the development of new types of burden tests to find adequately powered SV-phenotype associations. We will build a novel statistical pipeline that employs the latest genetic association concepts and incorporates SV impact assessments from Aim 2 to discover disease-associated SVs from the full group of samples across the various projects of the TOPMed program.

This systematic and comprehensive investigation of complex SVs will yield valuable new resources, including a reference catalogue of SV events from thousands of individuals, and a standard set of tools and pipelines for performing functional SV analysis and association of SVs with disease. In addition, we will identify novel associations between SVs and disease phenotypes, contributing to the broader TOPMed goal of understanding the fundamental biological processes that underlie heart, lung, blood and sleep disorders. Finally, we will develop and distribute the best SV detection, functional prioritization and association solution to members of TOPMed.