**SPECIFIC AIMS**

Structural variations (SVs), such as deletions, duplications, insertions, inversions and translocations, comprise a powerful class of phenotype-shaping genetic variation that is widespread in human genomes. SVs are structurally diverse, ranging from “simple” events to complex rearrangements. Investigating SVs, particularly complex SVs, not only provides for a much more comprehensive assessment of genetic variation in a given human being but could also hold the key to a deeper, more mechanistic understanding of common diseases of the heart, lung and blood—providing a compelling rationale for comprehensive SV discovery and analysis as part of the TOPMed program.

However, SV discovery remains a challenge for all current algorithms. Most currently available algorithms lack the specificity for analyzing SVs at nucleotide resolution; this capability is critical in the context of complex SVs. Moreover, complex SVs are disproportionately observed in non-coding regions of the genome, making functional interpretation and analysis challenging. Finally, it is likely that many of the high-impact SVs, particularly complex SVs, will be individually rare and thereby evade detection by population-based tools. Surmounting these issues will depend on novel computational methodologies for i) mining these complex datasets for SV discovery at high resolution and large scale, ii) functional interpretation of SV origin and phenotypic effects, and iii) accurate genotyping and association of specific (and rare) SVs with disease.

With the present application, we seek to develop methods that will discover the link between genomic structural variations and diseases of the heart, lung and blood through a comprehensive analysis of SVs. Our proposal brings together a team of pioneers with a proven record of collaboration with one another and others, and of innovation in the field of SV discovery and large-scale functional genome analysis. The group will establish a novel platform for discovering, validating and genotyping complex SV events from the hundreds of thousands of genomes being sequenced by the various projects of the TOPMed program. Using insightful methodologies based on data from the TOPMed program and public repositories such as ENCODE and elsewhere, we will functionally annotate the variants and perform association studies in a disease-specific context, genotyping a selection of complex SVs in the full cohort of individuals sequenced. This systematic and comprehensive investigation of complex SVs will yield valuable new resources for future investigations, including a reference catalogue of SV events from thousands of individuals and a standard set of tools and pipelines for performing such studies.

Towards these goals, we propose the following three Specific Aims:

**Aim 1. Build an integrative pipeline for large-scale discovery of complex structural variation.** We propose to build an integrated, smart and scalable pipeline of popular SV calling algorithms developed by our group and others, to discover all classes of SVs in a select large 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 anticipate that repertoire of SVs will impact both coding as well as non-coding regions; thus, methods for functional SV assessment need to be genome-wide. We propose to develop a framework to evaluate SVs impacting protein-coding genes, non-coding RNAs and non-coding regulatory regions that will account for the varied ways an SV can affect genomic elements. This framework will integrate conservation information, existing genomic annotations and epigenetic/transcriptomic datasets from the TOPMed program as well as public sources such as ENCODE to assign a ***Functional Impact score*** to each SV. We will also upweight the impact score of SVs overlapping elements with ubiquitous activity, high network connectivity (i.e., hubs) and strong allelic activity (i.e., functional sensitivity to variants).

**Aim 3. Association of structural variants with common and rare diseases.** We anticipate that many of the high-impact SVs will be relatively rare, necessitating the development of new burden tests to find adequately powered SV and phenotype associations. We will build a novel statistical pipeline that employs the latest association concepts and incorporates SV impact assessments from Aim 2 to discover disease-associated SVs from the full ~100,000 samples across the various projects of the TOPMed program. Building a reference database of complex SVs in individuals in Aim 1 will be essential to this undertaking.