RESEARCH STRATEGY

SIGNIFICANCE

Structural variations (SVs), such as deletions, duplications, insertions, inversions and translocations, are among the most significant determinants of human genetic diversity to have been discovered. SVs affect far more bases than single-nucleotide polymorphisms (SNPs) combined. SVs can markedly affect phenotype in many ways, including modification of open reading frames, production of alternatively spliced mRNAs, alterations of transcription factor (TF) binding sites and structural gains or losses within the regulatory regions. Consortium efforts such as the 1000 Genomes Project (1000GP) estimate that a typical genome contains 2.1–2.5 thousand SVs, affecting ~20 million bases, or ~5–6 times that of SNPs. Beyond "simple" SVs, there is a growing appreciation for "complex" SVs in human genomes, which vary considerably in their architecture and show complex patterns of rearrangements between distinct loci and/or even different chromosomes[1]. Through the 1000GP, we found that a large fraction of SV events have much higher breakpoint complexity than previously estimated—suggesting that complex SVs, like simple SVs, are also widespread in human genomes.

SVs are common, larger in size and more structurally diverse than single nucleotide variants (SNVs), so they are likely to profoundly shape the regulation of many human phenotypes and disease states. Investigating SVs, and particularly complex SVs, could therefore hold the key to a deeper, more mechanistic understanding of rare and common diseases. At present, most studies do not capture the spectrum of complex SVs present in genomes, so this complexity is not adequately accounted for in disease association studies. Furthermore, the functional impact of SVs, especially in noncoding regions, has not been investigated systematically. Surmounting these issues will depend on novel computational methodologies for 1) mining whole genome sequencing datasets for SV discovery at high resolution and large scale, 2) functionally interpreting their origins and phenotypic effects, and 3) establishing associations between specific SVs and disease.

Here, we propose to develop and apply novel methodologies to advance the overarching goals of the TOPMed program through computationally driven discovery, functional validation and characterization of disease-associated SVs. We will integrate novel and powerful tools for high-resolution SV discovery and use these to comprehensively profile all types of SVs, including complex SVs, from a large subset of the genomes being sequenced (Aim 1). To examine the functional impact of the identified SVs, we will integrate RNA-seq data and develop novel methodologies for functional annotation of variants and characterization of associated biological processes (Aim 2). Finally, we will scale up SV detection and analysis and genotype all SVs detected in Aim 1 across the ~100,000 samples of TOPMed, 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 platform of cloud-based pipelines for comprehensive, high-resolution and large-scale SV analysis. These will greatly enhance the ability of the TOPMed program to connect genetic variation to phenotypes of heart, lung, blood and sleep disorders.

Scientists participating in the proposed project are leaders in SV discovery and analysis. The three PIs, Charles Lee, Ph.D., Mark Gerstein, Ph.D. and Li Ding, Ph.D., have a history of productive scientific collaboration and bring complementary experience in SV detection (Lee), functional interpretation (Gerstein) and large-scale data analysis (all), particularly association analysis (all). Each also brings significant experience in leading (1000GP SV group, Lee; modENCODE AWG, Gerstein; ENCODE networks group, Gerstein; PsychENCODE AWG, Gerstein; exRNA AWG, Gerstein) and participating in (1000GP, Lee/Gerstein/Ding; ENCODE, Gerstein; ICGC, Gerstein/Ding; KBase, Gerstein; GSP (Genome Sequencing Program), Gerstein) large-scale sequencing consortia. Under Dr. Lee's leadership, the 1000GP SV project identified SV events in ~2,500 healthy genomes and helped define the methodologies for identifying and characterizing SVs from "lower depth" (~4X) whole genome sequencing (WGS) datasets.

INNOVATION

The originality of this proposal lies in the integration of cutting-edge computational methodologies—pioneered by the group—into a comprehensive, cloud-ready platform for novel SV discovery, characterization and association with common human diseases. The TOPMed Program will require high-resolution SV analyses that can be implemented at the immense scale required for adequately powered association analyses. Our proposed detection and genotyping strategy will meet the need for power and resolution for investigating association between SVs (that span a large size spectrum) and various phenotypes, surpassing previous standard approaches employed in current SV association studies. The key innovations of our approach lie in its characteristics of: **1)** *Scalability*: Our cutting-edge SV detection and integration tools will provide the

TOPMed Program with the capability to perform high-resolution classification of complex SVs, and identify well-powered genotype-phenotype associations in a disease context, across 100K genomes. **2)** *Integration*: Our approach will integrate identified SVs with RNA-seq data and other functional data from coding and non-coding (nc) regions of the genome to provide scores for functional impact. **3)** *Extended functionality*: Tools for mechanistic interpretation of SVs across different classes will allow us to make inferences about population structure and human adaptation and evolution. **4)** *Sensitivity*: Association tests that integrate weighting methods for various biological considerations, such as allele frequency and impact score, will enable a generalized linear model to capture subtle association signals often missed by conventional approaches. **This systematic survey of complex SVs will yield the largest reference database of validated SVs to date, together with an unparalleled system for high-dimensional, high-resolution studies of SV architecture and function in health and disease.**

RESEARCH STRATEGY Specific Aim 1. Build an integrative pipeline for large-scale discovery of complex structural variation.

Rationale. To drive the discovery phase of the program, we are currently working on *fusorSV* (manuscript in preparation, **Figure 1**), a framework developed by our group to discover SVs in hundreds of sequenced whole genomes. *fusorSV* takes a data mining approach to SV calling by incorporating knowledge of the strengths of various existing SV callers (discovered using a truth set), and uses this knowledge to perform discovery on a novel cohort of genomes using an ensemble approach. We will apply the fusorSV framework to a discovery cohort of individuals being sequenced by the Phase I/II TOPMed program projects. Using breakpoint assembly methods, we will perform *in silico* validation (**Figure 2)** of the SV events and use the assembled contigs to investigate the inherent complexity prevalent at breakpoints. **Ultimately, these studies will deliver the most comprehensive library of validated SVs discovered in humans and empower us to make novel biological inferences at the population level and in disease-specific contexts.**

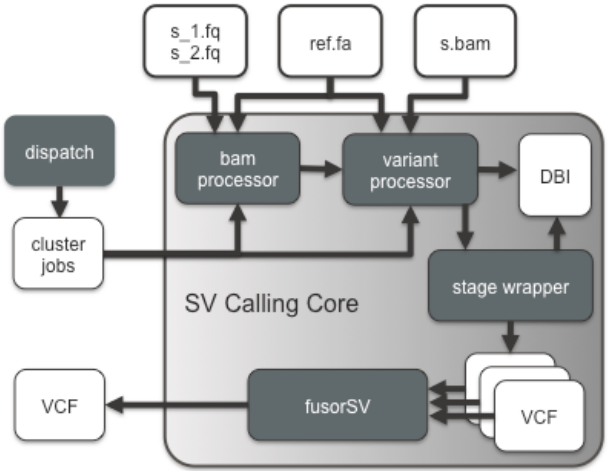

Figure 1. Structural Variation Engine that drives the various components of the fusorSV framework.

Preliminary data. *A toolbox of methods for structural variation discovery*. As part of the 1000GP SV project, we have provided the research community with an unprecedented set of germline SVs from more than 2,500 normal human genomes that have been sequenced at low depth and have developed a large toolbox of complementary tools and methods, including: ***(i) Read depth–based tools***. We developed CNVnator[2] for copy number variant (CNV) discovery and genotyping from individual and trio-sequencing datasets. It utilizes a mean-shift approach, GC correction and bandwidth partitioning to identify a wide range of CNV events. CNVnator can detect CNVs and provide genotype information on a population level, and also detects atypical CNVs including *de novo* and multi-allelic events. ***(ii) Paired end–based tools***. Meerkat[3], Hydra-Multi[4], PEMer[5] and BreakDancer[6] cluster abnormally mapped paired-end reads to identify loci with a signature for an SV event. Meerkat remaps soft clipped and unmapped reads to generate clusters to identify breakpoints. Pindel[7] utilizes a pattern-growth approach to detect large deletions and

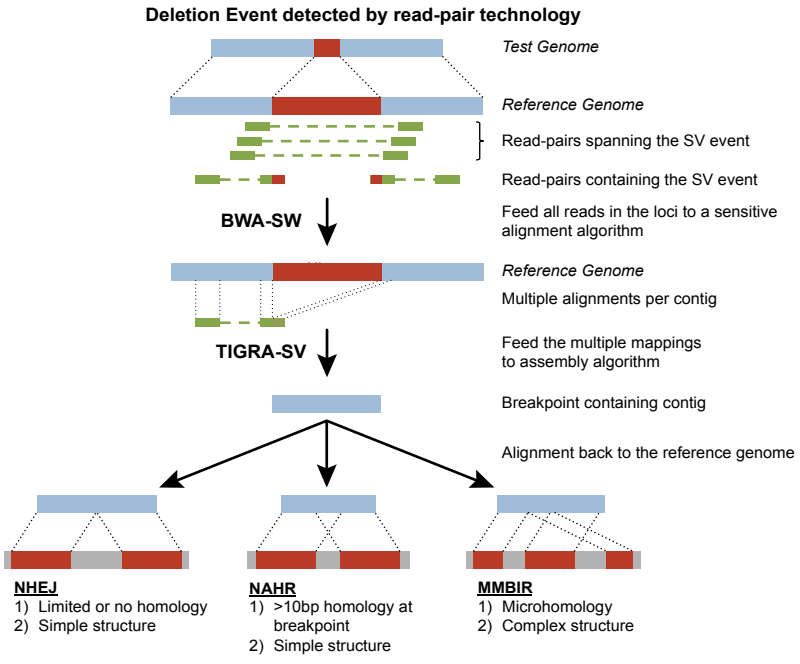

Figure 2. Breakpoint assembly for in silico validation. The top half of the figure shows a deletion SV event predicted by the readpairs spanning the event. All read pairs in the breakpoint locus are used for targeted *de novo* assembly and the resulting contig is aligned back to the genome.

insertions from WGS data. These methods have each already been successfully applied to hundreds of cancer genomes[3,8]. ***(iii) Split read alignment–based tools***. We have also developed SRM[9] and SRIC[10] for the high-resolution identification of SV events from WGS datasets. These tools specifically aim to provide base-pair resolution of breakpoints—an invaluable feature that enables functional interpretation of the biology of these SV events.

Breakpoint assembly tools for in silico validation. We also developed algorithms for identifying breakpoints at nucleotide resolution, thereby allowing us to validate SV breakpoints *"in silico"*. As previously described[8], we used assembly-based methods like SGA[11] or TIGRA-SV[12] for generating sequence contigs at breakpoints. Aligning these contigs back to the genome in the expected location and orientation validates the SV call (**Figure 2**). Using this method, we validated 64.8% of somatic breakpoints and 58.5% of germline control breakpoints[8]. We also developed AGE[13], which performs sequence alignment at regions flanking SVs while considering large deletion and insertion blocks, which cannot be handled by conventional sequence alignment algorithms.

Tools for complex event identification and assembly. It is now recognized that a large fraction (10–20%) of SV events are complex in nature[8,14]. We were one of the first groups to define the rules to identify complex rearrangements[8] from WGS datasets. Using these rules we comprehensively characterized complex SVs from a large cohort of TGCA WGS datasets[8] and validated them *in silico*.

Extensive complexity at structural variation breakpoints. As part of the 1000GP SV analysis team, we assessed the complexity of deletions where breakpoints had been sequenced and assembled. Consistent with the clustering analysis and the observed repeated rearrangement of duplication sites, 7.1% (1822) of these deletions intersected another deletion with different breakpoints. A larger fraction (16%) of assembled deletion sites had additional inserted sequence at deletion breakpoints. To further examine variant complexity, we grouped 1,651 deletions with at least 10 bp of additional DNA sequence between the original SV site boundaries into four classes (**Figure 3a**). The most common, *Ins with Dup and Del*, (N=501, 30%), exhibited a recognizable duplicated sequence interval within the inserted sequence. Not all SVs fit neatly into the classes depicted in Figure 3a, with 214 sites forming distinct patterns exhibiting increased breakpoint complexity. Within the 1000GP sample cohort, we also found that an appreciable fraction (80%) of inversions are complex (**Figure 3b**), likely involving DNA replication errors[15,16]. These results reveal the extensive complexity of SV breakpoints and highlight the importance of mining this complexity at fine resolution for interpreting SV biology.

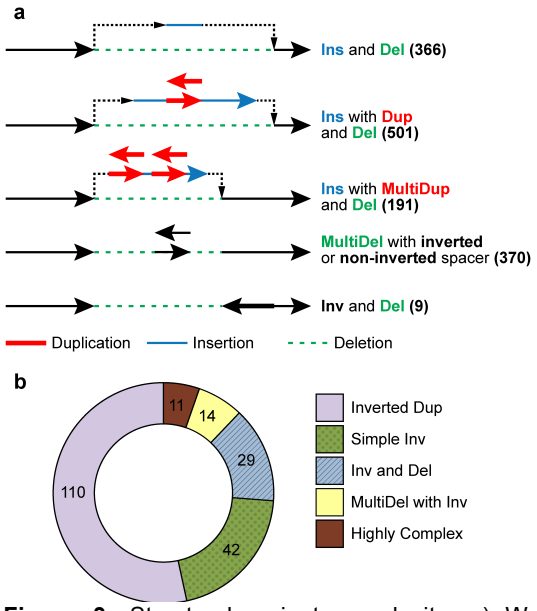

Figure 3. Structural variant complexity. a) We analyzed complexity of ~30K deletions from the 1000 GP phase 3 dataset, and characterized the events into several categories based on amount of complexity observed at the locus. b) A similar study of breakpoint complexity was performed for inversion events and revealed much higher levels of complexity than expected.

Ensemble approach to SV discovery. fusorSV (Figure 1, manuscript in preparation) is a framework that employs a data mining approach to integrating many complementary SV callers for analyzing very large cohorts of genomes. *fusorSV* allows for germline, somatic, and *de novo* SV analysis in the cloud or on traditional high-performance compute clusters. We took deep-coverage, PCR-Free WGS data from 27 samples sequenced by the 1000Genomes Project. Using the annotated SVs from the 1000GP Phase 3, we then performed k=3 cross fold validation on this cohort, wherein we built a model using 18 samples and applied the model to the other 9 samples for SV discovery *ab initio*. This step was repeated 1000 times with random selection for the learning samples and the test samples. **Figure 4** shows the performance of *fusorSV* as compared to some of the popular SV callers (BreakDancer[6], BreakSeq[17], cnMOPS[18], CNVnator[2], Delly[19], GenomeStrip[20], Hydra-Multi[4], Lumpy[21]) that were integrated using *fusorSV*. As can be seen, *fusorSV* outperforms all the SV callers by optimizing both precision and recall on the 1000GP Phase 3 callset. Even with a strict metric such as the Jaccard Similarity score[22], *fusorSV* outperforms all other SV callers for SV discovery in the test set. This Specific Aim will build on this framework and incorporate many other novel computer algorithms and improve performance.

Research Plan. We plan to develop new tools and extend the *fusorSV* framework to identify and classify SVs across WGS datasets from the various projects of the TOPMed program. The new and improved *fusorSV*, will deliver 1) integrated and comprehensive identification of a broad spectrum of SV types created by different molecular mechanisms; 2) compatibility with second- and third-generation-sequencing technologies and 3) breakpoint resolution identification based on TIGRA-SV (and other tools) and local assembly for *in silico* validation of the SV event.

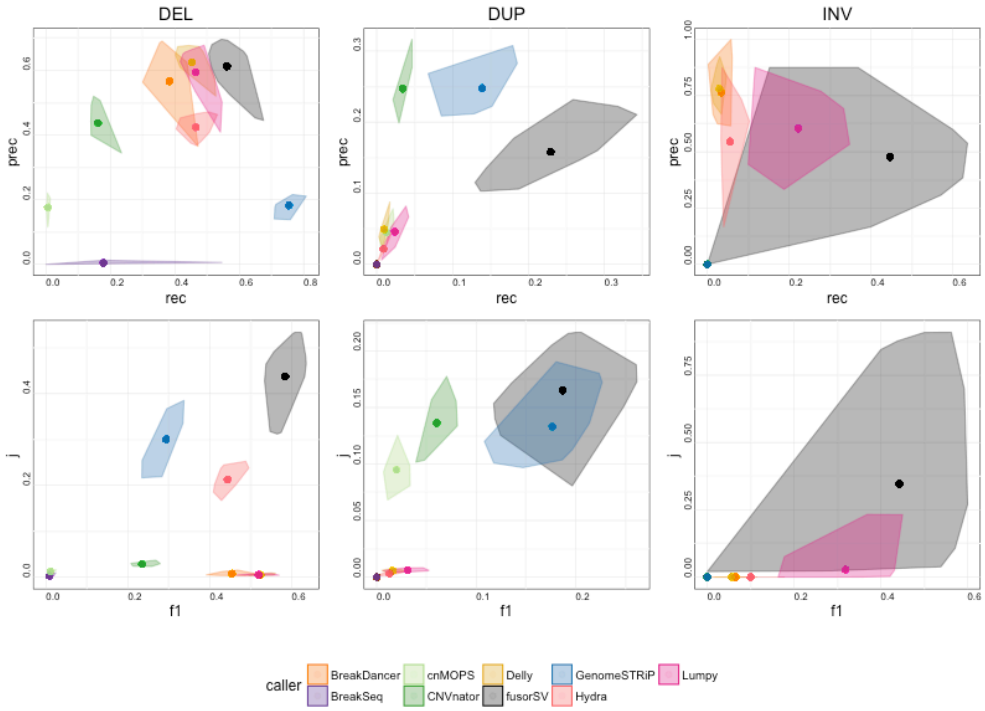

Figure 4. k=3 fusorSV cross fold validation using 1000 Genomes Phase 3 samples. The top 3 panels plot precision (y-axis) VS recall (x-axis). The bottom three panels plot the jaccard similarity score (y-axis) VS the F1 harmonic mean (x-axis)

Sample selection. Data storage and compute requirements preclude SV discovery on the whole TOPMed program. Based on our power calculations (Aim 3), we will select a discovery cohort of 10K individuals across Phase I/II projects for *de novo* SV calling. This will be important to assess the applicability and efficiency of our pipeline using datasets generated from different sites. We will prioritize sample selection based on availability of orthogonal datasets (e.g., RNA-Seq, Methyl-Seq etc) and phenotypic information (e.g., blood pressure, glucose levels, BMI, etc). Clearly, having additional genomic/phenotypic data would allow us to mine better biological inferences from the SV calls.

Pipeline for population-level structural variant discovery. During phase 3 of the 1000 GP SV project, we used an ensemble of nine algorithms for SV discovery. Individual call-sets were merged into a single release through a procedure that involved re-genotyping SV genomic loci using GenomeStrip[20] with an emphasis on genotype concordance for overlapping sites. The proposed *fusorSV* framework (**Figure 1**) for SV discovery will extend this work with the following salient features: 1) MySQL database–based sample tracking of data files through the pipeline; 2) Standard steps for quality control, duplicate removal and alignment for all selected samples; 3) An ensemble of SV-calling methods including CNVNator[2], cnMops[18], BreakDancer[6], Pindel[7], Hydra-Multi[4], Delly[19], BreakSeq2[17], Lumpy[21] and GenomeStrip[20]. This ensures that a particular algorithm does not bias the discovered SV set and increases our power to detect true SV events by asking for evidence by multiple methods; 4) Unified methods for SV genotyping and phasing using the lessons learnt from Phase 3 of the 1000GP[23]; 5) Validation for discovered set of SV sites using a library of known common variants and a targeted *de novo* assembly–based approach; 6) Complex SV identification using tools for assessing breakpoints at nucleotide resolution.

The SV calling will be performed in three phases:

*Phase 1—Calibration (Tasks 1,2 in **Figure 5**):* The pipeline will be developed using a machine-learning approach to calibrate and test the parameters of the different SV-calling methods. We will initially focus on 50 deep coverage "known truth" (KT) samples from the

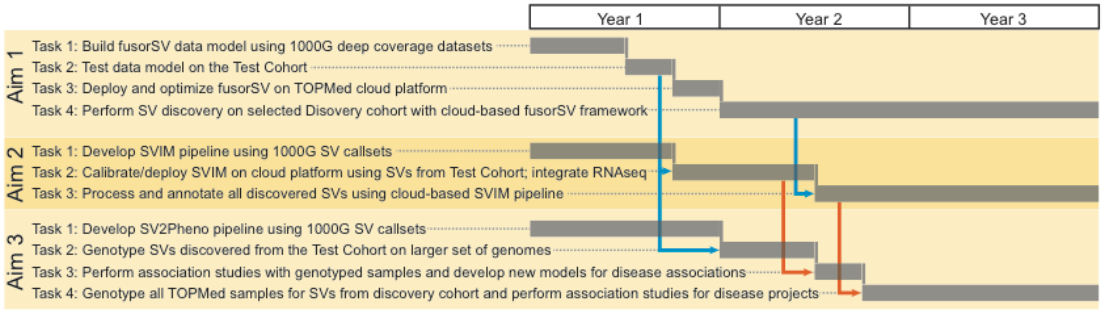

Figure 5. Project timeline. Arrows show flow of data between Aims. Aim 1 feeds both Aims 2 and 3 (blue arrows), and Aim 2 feeds Aim 3 (orange arrows).

1000GP SV Project[23], 100 "simulated truth" (ST) samples generated using WGSim (https://github.com/lh3/wgsim), and 200 test cohort (TC) samples (from the TCGA consortium). These datasets all contain some true-positive SVs and will be weighted in eventual determination of pipeline parameters depending on the level of confidence in the associated SV set (KT>ST>TC).

Phase 2—Optimization (Task 3): After calibrating our methods on the ST, KT and TC cohorts, we will expand the analysis to ~1% (~1,000) of individuals being sequenced within the TOPMed program. This cohort will be used to test for efficiency and eventual scale up in the next discovery phase. Based on the data access and compute strategies defined in TOPMed, we will explore parallelization where the tools already support this capability. The compute-intensive steps in the discovery pipeline that would be primary candidates for optimization are 1) genome alignment of raw reads, 2) clustering of aberrant reads, 3) SV validation using assembly and most importantly 4) SV integration.

Phase 3—Discovery (Task 4): The optimized system will be run on 10K of the proposed 100K individuals sequenced by the various projects. For sample selection we will prioritize projects that have RNA seq data available as well as well annotated phenotypic data for their cohort.

Calibration of method using known sites. Hundreds of sites across the human genome are polymorphic in a large fraction of the population[24,25]. Phase 3 of 1000 GP SV project[23] showed that a significant fraction of SVs (35%) occur at a high frequency in the population (VAF ≥ 0.2%). We will create a catalog of common copy number polymorphic sites across the genome and use them as validation sites for our SV-calling methods.

Validation of SV sites using in silico assembly–based methods. We demonstrated above that SVs can be validated *in silico* using targeted *de novo* assembly–based methods (TIGRA-SV and SGA). The same methodology will be integrated into the fusorSV framework and will be used to process every discovered SV site for validation.

Complex SV identification. We will use two methods for complex SV identification. The first[8] identifies SV clusters present in the same genomic region that have similar allele frequencies and copy number ratios. This will help select SVs that are part of the same complex SV event. The second method[23] involves inspecting the mapping patterns of various parts of the assembled contig at the SV site. This would allow us to identify mislabeled SVs and SVs with more complexity than annotated by the individual SV-calling methods.

Data access strategies: The JAX SV Cloud. Total storage of the discovery cohort is expected to require ~4 PB based on TCGA WGS statistics. To deal with the data footprint and computing requirements, we propose to develop the JAX SV Cloud, which will be available to all members of our teams. Our two-stage local and cloud approach is as follows:

i) The JAX local data center. In a traditional center, data are downloaded for analysis to local high-performance compute resources. JAX has extensive infrastructure, including an HPC cluster with 1856 cores and 2 PB of storage that will further expand over time (see Facilities and Resources). We can analyze the full discovery cohort by transient download and analysis of raw data with retention of only necessary results.

ii) Cloud-based data access model. However, it is expected that the TOPMed program will provide access to the data using a public cloud service provider. After initial method development and analysis, we plan to disseminate methods to the broader research community using the cloud paradigm decided by the TOPMed program. JAX is currently expanding capabilities in cloud-based data analysis to address issues including access to increased compute power, co-localization of novel and reference datasets and reproducibility of analysis pipelines. JAX staff have adapted multiple pipelines for the Amazon cloud and evaluated the suitability of Amazon archival storage for genomics datasets. Dr. Ding's group has been developing GenomeVIP, a secure, HIPAA-compliant, web-driven variant discovery and annotation platform through which multiple independent analysis tools can be applied to a given dataset. As it can call upon both local HPC and Amazon cloud resources, GenomeVIP is a tool that we may initially use to assist with variant discovery and to download results to local disks for subsequent analyses.

JAX is partnering and collaborating with commercial genomics cloud service providers (CSPs, Seven Bridges Genomics) to on several impactful projects and has recently recruited cloud computing experts as part of the Research IT department. These activities are independent of this U01 proposal. These efforts parallels that of the U01, namely to ensure methods developed at the data center will be stable and easily usable by the general research community.

Expected results. These studies will yield a comprehensive catalog of validated complex SVs from healthy and diseased individuals that lay the foundation for subsequent functional interpretation and association studies (Aims 2,3). They will also help answer questions about complex SV formation and population-level associations of SVs across multiple studies, thereby adding value to the TOPMed datasets. By making the fusorSV pipeline available as a community resource, and demonstrating the correctness and comprehensiveness of the SV results, we expect this work to propel future genome-level SV analyses for the entirety of the TOPMed program.

Pitfalls and alternative approaches. A major challenge for this study is the diversity of phenotype data that is being collected and of the variable availability of orthogonal data (genomic, transcriptomic, proteomic, etc.) across the various project cohorts being studied. In response, we will leverage the extensive experience of the team to handle complex datasets (see Prelim data section) and design *fusorSV* to robustly handle diverse and complex datasets of the type that might be generated by the TOPMed Program. Another potential pitfall comes from the current lack of defined cloud-based strategy by which the TOPMed Program will provide access to primary data. As a fallback plan, we can process samples by transient downloads / shipment of hard disks directly from the DCC and analyse the samples on our extensive local resources.

Specific Aim 2. Develop tools to analyze the functional impact of structural variants.

Rationale. There is still little known about the functional impact of SVs at a genome-wide level. SVs are disproportionately observed in the non-coding part of the genome; hence, a comprehensive assessment of the functional impact of SVs will likely require the integration of large-scale data resources such as ENCODE, 1000GP and GTEx. To functionally prioritize SVs in preparation for disease association studies, we propose to create **SV Impact (SVIM)**, a new analysis tool that integrates myriad datasets- including existing annotations, allelic activity from RNA-seq, and eQTLs from RNA-seq.

Preliminary data. *Tools for assessing functional impact of genomic variation in genes and pseudogenes.* We developed Variant Annotation Tool (VAT) to annotate the impact of protein sequence mutations. VAT provides transcript-specific annotations of point mutations and indels according to synonymous, missense, nonsense or splice-site-disrupting changes[26]. We observed that genes tolerant of loss-of-function (LoF) mutations are under the weakest selection. In 1000GP Phase 3, we found that a typical genome contains ~150 LoF variants and discovered significant depletion of SVs (including deletions, duplications, inversions and multiallelic CNVs) in the coding sequences, untranslated regions and introns of genes compared to a random background model, implying strong purifying selection.

Tools for evaluating functional impact of variation in non-coding (nc) RNAs and regulatory regions. We developed tools to specifically analyze ncRNAs. Our incRNA pipeline combines sequence, structural and expression features to classify newly discovered transcriptionally active regions into RNA biotypes such as miRNA, snRNA, tRNA and rRNA[27]. Our ncVar pipeline further analyzes genetic variants across biotypes and subregions of ncRNAs, e.g., showing that miRNAs with more predicted targets show higher sensitivity to mutation in the human population[28].

To better understand nc regulatory regions, we developed tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. PeakSeq identifies regions bound by TFs and chemically modified histones[29,30]; it has been widely used in consortium projects such as ENCODE[29,31]. The second generation of PeakSeq is a newly developed tool that uses multiscale decomposition to help identify enriched regions in cases where strict peaks are not apparent and robustly calls both broad and punctate peaks[30]. Peak calls and ChIP-Seq signal data can also be used to model gene expression and annotate target genes. We have developed methods that use both supervised and unsupervised machine-learning techniques to identify these regulatory regions (such as enhancers) and predict gene expression from ChIP-Seq data[32-35]. To investigate the evolutionary importance of these regions, we have analyzed patterns of single nucleotide variation within functional nc regions, along with their coding targets[28 35,36]. We used metrics such as diversity and fraction of rare variants to characterize selection pressure on various classes and subclasses of functional annotations[28]. We have also defined variants that are disruptive to a TF-binding motif in a regulatory region[31].

Tools for helping annotate functional impact based on network. We found that functionally significant and highly conserved genes tend to be more central in various biological networks[37] and are positioned at the top of regulatory networks[36]. Further studies showed relationships between selection and protein network topology (e.g., quantifying selection in hubs relative to proteins on the network periphery[37,38]). Incorporating multiple network and evolutionary properties, we developed NetSNP[37] to quantify the indispensability of genes. This method shows strong potential for interpreting the impact of variants involved in Mendelian diseases and in

complex disorders probed by GWAS. We constructed regulatory networks for data from the ENCODE and modENCODE projects, identifying functional modules and network hierarchy[36]. To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM[39]).

FunSeq: Tools for integrated functional prioritization. We recently developed a prioritization pipeline called FunSeq[40,41] that identifies annotations under strong selective pressure as determined using genomes from many individuals from diverse populations. FunSeq links each nc mutations to target genes and prioritizes based on scaled network connectivity. FunSeq identifies deleterious variants in many nc functional elements, including TF binding sites, enhancer elements and regions of open chromatin corresponding to DNase I hypersensitive sites, and detects their disruptiveness in TF-binding sites (both LoF and gain-of-function events).

Mutational mechanisms of structural variants. The sequence content of SVs, especially around breakpoints, carries important information about origin and functional impact. Using datasets from 1000GP, we studied the distinct features of SVs originating from different mechanisms[40,42]. We performed SV mechanism annotations for the 1000GP Phase 3 deletions using BreakSeq[17], categorizing 29,774 deletions by their creation mechanisms. Among these, NHR proved to be the most prevalent mechanism (~73% of all categorized deletions)[23]. These results inform us on the molecular mechanisms underlying SV formation and also indicate differences in functional impacts of different SV types.

Tools for uniform processing of RNA-seq data. We have considerable expertise in analyzing RNA-Seq data, including experience in developing and setting up pipelines for the processing of RNA-seq data; specially for long RNA-seq data for ENCODE, long and short RNA-seq data for the PsychENCODE[43] and Brainspan project as well as a custom pipeline developed for the analysis of small exRNA-seq data for the Extracellular RNA Communication Consortium (ERCC). We have already developed an efficient in-house data processing workflow for RNA-seq data that includes data organization, format conversion, and quality assessment. RSeqTools[44] is a modular tool developed for the processing of RNA-seq data and generating either transcript, gene or exon level quantifications. We also developed IQSeq[45] which calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-seq data using a fast algorithm based on the Fisher information matrix. Another tool we developed called FusionSeq[46] was to detect fusion transcript in RNA-seq data, which can be important biomarker for diseases such as various types of cancer and neurological diseases.

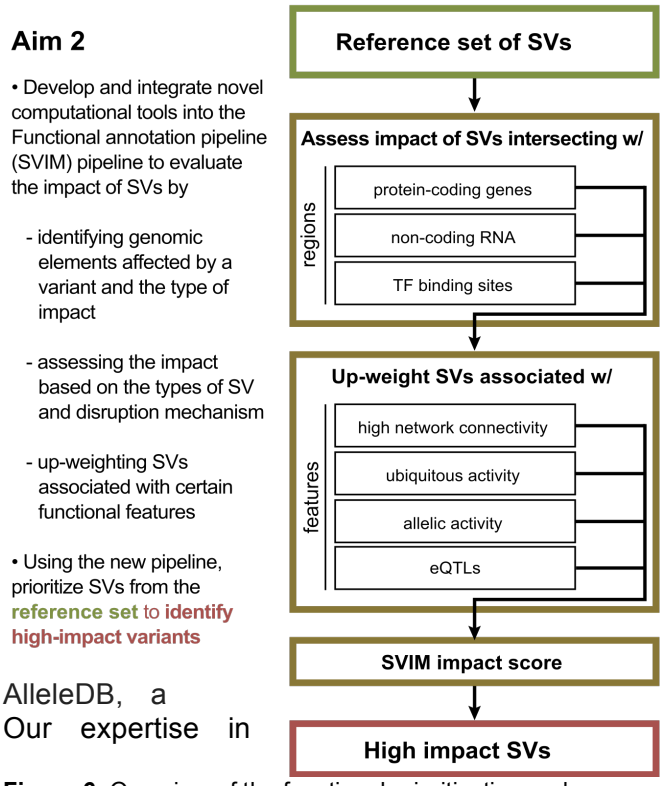

Figure 6. Overview of the functional prioritization and annotation pipeline

AlleleDB, a Our expertise in

Tools for allele activity and eQTL detection. We have also developed tools specifically for linking gene expression variation to genotype, including our Allele-Seq pipeline, which quantifies allele-specific gene expression by mapping reads onto a diploid personal genome built from called genetic variants, including SNPs, short indels, and structural variants [47]. We recently applied this pipeline on a population scale to RNA-Seq data from the 1000 Genomes Project, and used this analysis to create database of genomic regions with high allelic activity[48]. eQTLs is demonstrated in our novel study on successfully utilizing expression-variant correlations to construct predicted genotypes. These predicted genotypes were then matched with known genotypes from a given dataset in order to demonstrate how the information security of the given dataset may be compromised[49].

Research plan. To enable identification of SVs with high functional impact, we will extend FunSeq/FunSeq2 within a new pipeline called SVIM (Structural Variation IMpact)(**Figure 6)**. We will evaluate the impact score for each SV, taking into account the functional annotation of the affected genomic region and the fraction of functional elements (i.e., genes, ncRNAs, nc regulatory elements). We will also upweight SVs based on

ubiquitous activity, allelic activity and eQTLs. The impact score will also depend upon SV type (i.e., deletion, duplication, inversion or translocation).

For a given SV belonging to a particular SV type, we will use break point resolution coordinates to estimate the fraction of bases overlapping functional elements. Based on this fraction, we will categorize SVs into three classes (touch, cut, and engulf). Each overlapping class will have a different weight ($F_{svtype,\ class}$). We will divide genomic elements into three categories (coding region, nc region, TF binding site) and assign relative scores to them (S_{coding}, $S_{non\text{-}coding}$, S_{TFBS}), which will vary for different SV types. Relative scores F and S will be defined for class and functional elements analogous to the FunSeq2 tool[40].

$IS_{orig} = \sum_i \left(F_{j,k} \times S_{j,i} \times \delta_i\right) \times \prod g_l$; $IS_{norm} = \frac{IS_{orig} - \overline{IS_{random}}}{\sigma_{random}}$, where i is a functional element $\in \{protein\ coding, noncoding\ RNA, nonconding\ regulatory, allelic\ activity, eQTL\}$; k is a overlapping classification $\in \{cut(\ 0.1 \leq f < 0.8), touch(f\ < 0.1), engulf(\ f \geq 0.8)\}$, and f is the fraction of functional element overlapping the SV; j is the type of SV ; $\delta \in \{0,1\}$; and l is a feature $\in \{connectivity, ubiquitous\ activity, allelic\ activity, eQTLs\}$;

SVs will be assigned an impact score by taking the sum over the product between weights of overlapping classes and scores of overlapping functional elements. The score (IS_{orig}) will also be upweighted based on activity of the affected region. The upweight factor is comprised of the product of four factors: i.e., allelic activity, eQTLs, network connectivity and ubiquitous activity. Significance level of an Impact score (IS_{orig}) will be estimated by running 1,000 Monte Carlo simulations generated by randomly shuffling the location of SVs.

Evaluating effect of structural variants on protein-coding genes. We will further develop a protein-coding module for SVIM to substantially expand the analysis of loss of function (LoF) variants with mis-mapping, functional, evolutionary and network features. We will first identify LoFs due to whole gene deletion, as well as putative LoF-causing mutations as those that induce premature stop codons, frameshifted open reading frames, or that we predict to produce truncated proteins due to deletion of RNA splice sites or either predicted or verified changes in splicing pattern from RNA-Seq data (see above). We will quantify the confidence of these LoFs using features such as whether they are in highly duplicated regions and the number of paralogs. For functional features, we will incorporate protein structures. For evolutionary properties, we will quantify the conservation of LoF variants, as well as truncated sequences. For network features, we will quantify the distance between genes with LoF variants and known disease-causing genes.

Prioritizing non-coding transcripts from structural variant data. To prioritize the effects of SVs in ncRNAs, we will focus on overlaps with regulatory elements and other functional regions. To perform this analysis, we will define categories of RNA regions that display human population-level conservation, and combine these features to generate RNA element scores. We will mine RNA interactions between proteins (e.g., CLIP-Seq) and miRNAs (e.g., TargetScan) to create a compendium of biochemical interactions with RNA[50-54]. We will further investigate RNA secondary structure, looking for structured regions that are highly sensitive to mutation. For these regions, we will assess deleteriousness of mutations by differences in predicted free energy or structure ensembles[55] relative to wild type. We have found annotations of all of the above types—biochemical interactions, regulatory motifs, and structured regions—that are enriched for rare variants in the human population and will use these sensitive RNA regions to score and prioritize potential deleterious SVs in ncRNA. Large SVs will ultimately be scored based on the highest scoring subregion disrupted (or created) by the SV.

Prioritizing non-coding regulatory elements from structural variant data. Unlike protein-coding genes and ncRNAs, TF binding motifs are relatively small in size. Thus, we are going to analyze duplications that occur close to these motifs and analyze where these duplications lead to the breakage of existing or creation of new motifs. In the prioritization scheme, we will also penalize changes in distance between motifs and newly created motifs if they occur close to an existing TF motif. We will use TF binding nc elements by leveraging better enhancer definitions provided by the Epigenome Roadmap[56-58] and ENCODE and also include new datasets.

Further variant prioritization based on networks, tissue specificity, eQTLs and allelic activity. After performing annotation-based assessment of identified SVs, the following functional features will be used for prioritization.
i) Network connectivity. We will examine the network topological properties of the genomic elements affected by identified SVs. Variants disrupting regulatory elements with high connectivity—network hubs and bottlenecks—will be upweighted based on their scaled centrality scores.

ii) Ubiquitous activity. We will evaluate the impact of SVs in an epigenetic context to identify tissue-specific phenotypic effects that are strongly influenced by SVs. We will prioritize SVs impacting genes, ncRNAs, and TF binding sites active in multiple tissues.
iii) Allelic activity. We will use our existing AlleleSeq pipeline to annotate the transcripts produced at SV regions [47]. We will use this tool to create personal diploid genomes for each TopMed individual, and then will adapt our pipeline to perform RNA-Seq quantification specifically at SV regions. We will prioritize SVs that lead to strongly allelic expression. We will also prioritize SVs that overlap our database of strongly allelic regions throughout the genome, based on AlleleDB, our resource of such regions identified through allele-specific RNA-Seq analysis from over 300 individuals generated by the GEUVADIS consortium[48].

iv) eQTL association. We will link SVs to the genes that they affect by performing genome-wide searches for eQTLs. Relative to SNVs, large SVs may be more manageable candidates in the search for distal eQTLs. We will use a framework similar to published earlier[49] in the search for SV-induced eQTLs. SV-induced eQTLs will be identified by performing genome-wide searches for patterns in which the presence or absence of the SVs (from Aim 1) strongly correlate with the expression levels of a battery of genes throughout the genome. Specifically, we will use Matrix eQTL for eQTL identification [59]. We will perform multiple testing correction and will filter the list of putative eQTLs in order to achieve a false discovery rate of less than 5%. The SV-gene expression correlations reported by Matrix eQTL will be used as the strength-of-association measures between expression levels and genotypes. Of particular interest will be those genes previously implicated in disease-associated pathways and network modules. SV-induced eQTLs with strong expression correlations that are associated with central network elements and known disease-associated genes will be upweighted.

Expected results. We expect that SVIM, a new software solution to estimate the impact scores of the SVs produced in Aim 1, will yield a prioritized set of SVs in Aim 2 that we can forward to Aim 3 (genotype and association) for further classification of their association to disease or a specific phenotype. We plan to make the prioritization results broadly available; therefore, SVIM will incorporate the impact score into a standard Variant Call Format (VCF). SVIM will be cloud-ready and will be available to the TopMed consortium through a Docker image and a Common Workflow Language (CWL) file.

Pitfalls and alternative approaches. We anticipate that the greatest pitfalls are (i) possibly an overwhelming number of SV discovered in Aim 1 and (ii) the lack of standard format and increasing number and updates of annotation datasets. In order to overcome (i), we plan to gradually process the results into specific type of SVs. SVIM will also be based on the data context to optimally prioritize from WGS datasets. The overall modularization offers a flexible framework for users to incorporate the ever-increasing amounts of genomic data to both rebuild the underlying data context and prioritize case-specific variants. In order to overcome pitfall (ii) we will make great efforts to make SVIM computationally efficient and able to support the large-scale computing proposed for this aim. To build the data context, we will standardize large-scale publicly available data resources, such as SVs from the 1000 GP[60], conservation data from Bejerano *et al.*[61] and Cooper *et al.*[62], functional genomics data from ENCODE[31] and Roadmap Epigenomics Mapping Consortium[63].

Specific Aim 3. Scaling up to 100K samples and associating SVs with common and rare diseases.
Rationale. Since many high-impact SVs are expected to be relatively rare, such that conventional association tools cannot readily and robustly handle them. Therefore to discover important SVs, we must develop a new association pipeline suitable for finding important SV-phenotype associations. We anticipate that building a reference database of complex structural variants in healthy individuals (Aim 1) will be essential for this goal.

Preliminary Results. *Power analysis for sample selection and association.* An important aspect will be selecting a subset of the 100K samples projected to be sequenced for full SV analysis. This discovery cohort will furnish the prototype events that will subsequently be studied in the full population by genotyping the entire sample set. Total analysis cost (e.g. downloading, storage, compute time, manual review) must be balanced against the discovery probability for events having the lowest population minor allele frequency (MAF) we wish to include. There is no general theory of discovery power currently used in SV algorithms, so we extended an existing statistical model of coverage[64] to estimate the discovery sample size. Bernoulli probabilities for two standard SV discovery methods, split reads and discordant read pairs, can be derived using probability theory considering read length, average and variance of insert length, SV length, etc. and subsequent incorporation of a detection rule, e.g. "≥3 split or discordant reads". Detection in each sample is binomial in the number of observations and discovery within sample set is likewise binomial in the detection and MAF probabilities.

Anticipated parameters for the WGS data to be generated for this project are 30X coverage per genome, average insert size of 400bp-600bp (20% coefficient of variation), 150bp reads, event detection based on ≥3 split reads or ≥5 discordant read pairs, and observation in at least 3 samples to constitute "discovery". The model predicts that split-read detection will predominate for simple SVs, as well as for complex events in which one sequence is replaced by another. Because split-reads depend only upon local alignment, power is essentially independent of the size of events (unlike for discordant read pairs), meaning it is primarily a function of sample size and MAF. **Figure 7a** shows power at MAF ≥ 0.1% is essentially 100% for 10K samples. It drops rapidly for lower MAFs, whose events are unlikely to be discovered in this study. Mosaicism is a potentially confounding factor, for example in blood samples where an event is not present in all cells. **Figure 7b** shows that power is not significantly impacted for 10K samples until mosaicism is quite significant.

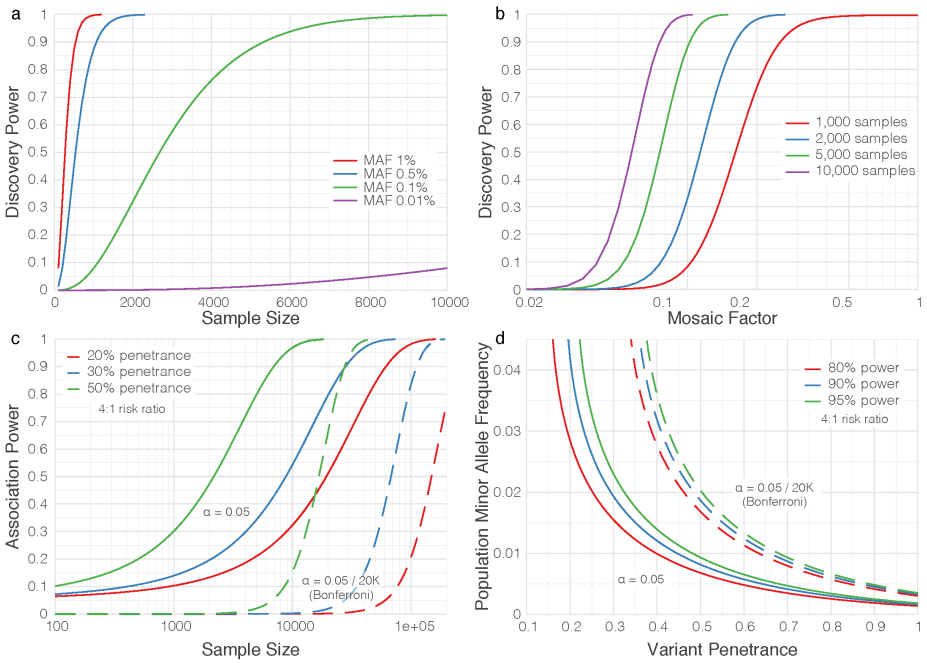

Figure 7. Power analysis for sample selection and association. a) Power vs sample size for selected MAFs from 0.01% to 1%. Events are assumed heterozygous and completely represented in the sample (no mosaicism). Curves are universal in that simple insertions and deletions, as well as complex indels, collapse and power is independent of indel size, since the "split reads" discovery mode dominates. b) Power vs "mosaic factor" (unity meaning event present in all cells; 0.5 meaning event present in half the cells, etc.) for selected samples sizes from 1K to 10K. All data plotted at 1% MAF. Split-read discovery again dominates and curves are universal. c) Association power for 10 collapsed variants (even numbers of cases and controls), each of 1% MAF and penetrance from 1% to 50%, at both single gene (α = 5%) and Bonferroni-corrected for 20K genes, as well as a 4:1 risk ratio for the Li and Leal (2008) collapsing strategy. d) Curves of constant power for 10K cases/10K controls, with other parameters the same as in c).

The second aspect of "power" is variant-disease association. The issues are well-known[65], enabling the following "baseline" estimates of association power. General consensus[65] recommends "collapsing" variants for low MAF in order to aggregate effects for increasing power. Analysis of the widely-used Li & Leal method for 10 collapsed variants at 4:1 risk ratio (**Figure 7c**) shows that groupings of 1% MAF variants having high (~50%) penetrance will require 20K-30K samples for 80% power when Bonferroni-corrected. Power drops rapidly for lower MAF, penetrance, risk ratio, and sample size. Although it is not yet known how the 100K samples will be divided over various studies, it is instructive to examine the scenario of 10K cases/10K controls (**Figure 7d**). Variants around 2% MAF should have ≥90% association power for penetrances ≥50%, while variants regardless of MAF having penetrances <25% will likely remain ambiguous, as will variants from phenotypes having substantially smaller sample allotments. It is likely we will discover more variants than those for which solid associations can be established.

Association pipeline implementation and experience in discovering significant associations. We have developed a prototype pipeline incorporating extensive sample and variant level quality control (e.g, coverage, variant frequency and distribution), population stratification, pedigree segregation, etc. for population/family-based association analysis. It supports popular aggregation tests, including burden tests such as the Combined Multivariate Collapsing (CMC)[65], Exclusive Frequency Test (EFT)[66], Total Frequency Test (TFT)[66], and Cohort Allele Sum Test (CAST)[67], and variant component tests such as the Sequence Kernel Association Test (SKAT)[68]. We have already used it to discover associations by tailoring it to hypothesized genetic architectures of individual diseases. For example, assuming tumor suppressors are enriched for rare deleterious truncations, we grouped events by gene and used TFT to associate 13 genes with germline susceptibility in a >4,000 case cancer cohort[69].

Research Plan. SVs are characterized by size, type, penetrance, and multiple alleles. We plan to genotype the top half of high-impact SVs detected in 10K discovery samples (**Aim 1**) across all ~100K samples to be sequenced by TOPMed centers to obtain sufficient statistical power for genotype-phenotype association. A

critical step for association analysis of SVs is meaningful classification/annotation. By building on infrastructure and tools mentioned above, we will develop a new pipeline called "**SV2Pheno**" to infer SV-phenotype associations (**Fig. 8**). It will use the impact scores for each SV (**Aim 2**) for integrated analysis of SNVs, indels, and SV.

Genotyping of SVs detected in the discovery set across the entire sample set. Genotyping and annotation of discovered SVs in the whole population will allow accurate determination of prevalence and allele frequencies and, importantly, increase association analysis power. This process will use **BreakSeq**[17] to build a library of validated and assembled SV breakpoints for genotyping individual genomes. For imprecise SVs, a combined read-pair/read-depth approach using **GenomeStrip**[70] will do population level genotyping. Conventional genotyping involves assembly of both reference and alternate sequence contigs, which are used as targets for mapping all reads present in the sample. However, given the large expected data footprint for the full sample set, the traditional "*bring data to the computing tools"* approach will be upended to "*bring compute tools to the data"*. We shall build on tools such as Sambamba (bam slicer function)[71], TIGRA-SV assembler and Pindel. This will reduce the footprint to a fraction of the original and enable the methods to work in the cloud and access data over a secure network.

Develop SV2Pheno pipeline including improved burden tests considering impact score and annotation classification of various complex structure variants. We envision substantial extension of this pipeline in two major ways to address the ambitious goals of this proposal: 1) We plan to hybridize the pipeline with more recent methods that better account for non-contributing variants[72]. Likewise, annotation and functional prediction can help identify irrelevant variants, which can subsequently be removed from analysis. The pipeline will also process the information from the ENCODE & Epigenetics Roadmap analysis mentioned in **Aim 2**. 2) Variants are known to be associated with various diseases[73-75], but almost certainly contribute non-uniformly; assigning appropriate weights will be necessary to wring-out maximum power. Aggregation tests can be expressed in general by the linear regression equation $Y = \alpha + \beta \cdot \Sigma w_i g_i + \varepsilon$, where (left-to-right) is observed trait, intercept, collective effect coefficient, weight of variant *i*, tally of variant *i* (0, 1, or 2), and normally distributed error residual. Assignment of weights will be based on a novel combination of four considerations: the Madsen-Browning equation[76] to account for allele frequency, consideration of "direction" (negative association) using e.g. aspects of the Pan-Shen approach[77], incorporation of our impact score (**Aim 2**) to account for biological strength, and RNA-seq data. The last aspect will weight expression impact, but must be implemented carefully because of variations in sample quality. Here, we will apply the method of Liu et al[78], which essentially adds an extra calculational adjustment to modulate contribution of higher-variability samples. In principle, this more sophisticated approach should capture signals that have been too subtle for earlier-generation tests[79].

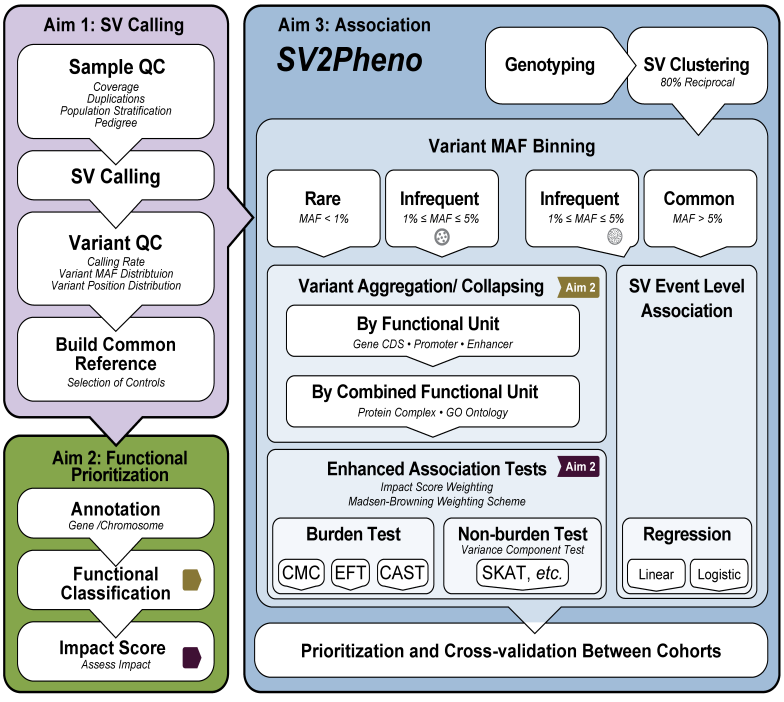

Figure 8. SV2Pheno Association Analysis Pipeline. The overall work flow includes QC, population stratification from Aim 1, functional classification and impact score generation from Aim 2 and single event test and burden analysis from Aim 3.

We are mindful that controls for each association analysis should be carefully matched with cases; paying close attention to population structure, sample coverage, etc. When sample size is fixed, an even case-control split offers maximal power. However, it is likely that the TOPMed program will furnish potentially many more controls and this increases power. For such diseases, we will check the available literature for any known underlying genetic commonalities and choose extra controls in light of relevant covariates (e.g. age or smoking status). Since we anticipate that a high fraction of SVs will reside in non-coding regions, we will aggregate variants using a hierarchical approach based on three levels:

Level 1. *Prototypical Event level association analysis.* As the precise genomic region for a given SV may vary across samples, we will

represent each set of similar SV events as a single prototypical SV event. The criterion constituting such events is given by the "80% reciprocal overlap" rule[12]. For large insertions and inter-chromosomal translations, we will require the breakpoints to be within 1kb of one another. We will then assess the significance of the associations using impact scores generated in Aim 2.

Level 2. *Functional Unit (Gene CDS/promoter/enhancer) level association analysis.* We annotate the prototypical SV events from Level 1 to identify any specific transcriptional regions (e.g., exons/CDS and cis-regulatory elements such as insulators, enhancers, and promoters) and gene(s). SVs in a given gene will be grouped as a single, effective functional unit based on annotation from **Aim 2** (**Figure 8**). We will then perform an association analysis on these functional units. In cases where multiple SV events may be affiliated with a given functional unit, we will develop a weighting scheme to combine the impact scores of the contributing SVs. This approach may reveal novel connections between non-coding functional regions and phenotypes.

Level 3. *Combined Functional Unit level analysis.* We will annotate the functional units in the previous step to identify known affiliated higher-order units (e.g., protein complexes and gene pathways) by recruiting various resources, including databases relating to gene-phenotype relationships (e.g., OMIM), gene pathways (e.g., KEGG, Reactome), gene ontology (e.g., GO database). The SVs affecting a given higher-order unit will be grouped as a single super-unit. We will again perform association analysis, considering the SV impact scores (**Aim 2**). This approach has the potential to discover novel combinations of SV-containing functional units.

We will apply this tiered approach and association analysis (**Figure 8)** to analyze all genotyped samples passing our extensive coverage and variant calling QC from various cohorts to identify promising candidate SVs associated with specific phenotype.

Integrate various types of variants for association analysis. The most powerful analysis will come by combining information from SNVs, indels, and SVs for association analysis. Traditionally, weights in burden tests account for variants with different MAFs, but favoring those having lower MAFs[68,76]. Bioinformatic information, such as PolyPhen scores for SNVs, and SV impact scores from Aim 2 will inform these weights. To the best of our knowledge, no previous approaches have aggregated variants of different types. Here, we propose two methods for such integration: 1) We hypothesize that SVs would have stronger functional impacts than missense SNVs, on average, and we will develop a weighing scheme based on the size and genetic architecture of various variant types using the framework of previous weighting schemes. SNV/indel/SV will be jointly calculated in a single burden analysis; 2) We hypothesize that alterations from functional regions, regardless of size, contribute to phenotype. Therefore, alternatively, we plan use SNV/indel and SV for independent burden analyses and combine the P-values from these independent tests.

Association between SNVs/indels and # of SVs. Under the null hypothesis that variation occurs randomly, it should be possible to correlate the numbers of SNVs/indels versus number of SVs, the slope being indicative of differences in rates of occurrence, and also to check such correlation against established rates. We will perform association analysis for individual outlier cases in which SV census is significantly lower or higher than expected. It is possible that such outliers might harbor common germline alterations leading to genomic instability by affecting DNA repair pathways.

Expected results. This aim will culminate in the **SV2Pheno** association pipeline and its tools for systematically discovering SVs associated with specific phenotype/disease. We expect to have increased statistical power to discover rare, novel SVs associated with phenotypes previously missed due to smaller sample size. We further anticipate revealing genetic changes associated with increased frequency of SVs genome-wide. The initial version of **SV2Pheno** will be distributed for broader community use and cloud distribution.

Pitfalls and alternative approaches. Our preliminary analysis indicates that we are well powered to detect SVs with MAFs around 0.5% to 1% using 10,000 cases. Although it is very likely that we will discover more SVs than we can establish associations for (discussed above), there are still some issues of selection. There are several strategies for selecting datasets for initial discovery: 1) from one homogenous cohort; 2) from one CCDG center across multiple cohorts; 3) from multiple cohorts generated by multiple TOPMed centers. Regardless of choice, we will maintain high standards regarding coverage, read length, insert size, mapping rate, % mismatch etc. to ensure accurate, representative detection of SVs across populations. To reduce the number of hypotheses to be tested, we can alternatively focus on SVs from regions indicated to have association with phenotype from the study of SNV/indel. The weighting methods discussed above for may require tuning and we will use known disease associated SVs as positive controls for the calibration.