**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 highly structurally diverse, ranging from these “simple” events to complex rearrangements between loci and/or chromosomes. Of those that have been associated with disease, they have proven to be highly impactful forms of genetic variation. Investigating SVs, particularly complex SVs, could therefore hold the key to a deeper, more mechanistic understanding of common diseases—providing a compelling rationale for prioritizing SV discovery and analysis among the analysis centers of the Genome Sequencing Program (GSP). However, a number of features of SVs render them challenging to study using current analytical approaches.

Owing to their structural diversity, SVs are often not adequately covered by current discovery methods, suggesting that a significant proportion with relevance to disease remain to be identified. Most available algorithms lack the specificity for analyzing SVs at nucleotide resolution; this capability is critical in the context of complex SVs, in which the sequences around breakpoints carry important mechanistic information. 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 conventional association 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) associating specific (and rare) SVs with disease.

With the present application, we seek to establish ***The Jackson Laboratory Center for Structural Variation Analysis (JAX CSVA)*** to facilitate the overarching goals of the GSP through comprehensive analysis of SVs. Led by Charles Lee, Ph.D., the JAX CSVA brings together a team of pioneers with a proven record of collaboration and innovation in the field of SV discovery and large-scale functional genome analysis. The JAX CSVA will establish a novel platform for discovering, validating and genotyping complex SV events from the thousands of genomes being sequenced by the data-producing centers of the GSP. Using cutting-edge methodologies, 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 in the GSP. The investigator-led component of the Center will focus on identifying SVs and their functional impact on disease biology, through which program-driven goals, i.e., methods for defining metrics and common controls for future common disease studies, will be achieved. 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. The JAX CSVA will also highlight the prevalence and importance of complex SVs in diseases and provide functional insight into complex SV formation and its effects. Towards these goals, we propose three Specific Aims:

**Aim 1. Build an integrative pipeline for large-scale discovery of complex structural variation.** We propose to build an integrated, efficient and scalable pipeline of tools developed by our group to discover all classes of SVs in a selected large cohort of individuals being sequenced as part of the GSP. 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 breakpoints. These studies will deliver the largest library of validated SVs discovered in humans and will allow us to make novel biological inferences.

**Aim 2. Develop tools to analyze the functional impact of SVs.** We anticipate that most SVs will impact 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 a SV can affect genomic elements. This framework will integrate conservation information, existing genomic annotations and epigenetic/transcriptomic datasets from sources such as ENCODE. 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 200K samples of the GSP. Results from this aim will extend the results obtained from a decade of GWAS studies. Building a reference database of complex structural variants in healthy individuals in Aim 1 will be essential to this undertaking.

**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; thus, they can markedly affect phenotype in many ways, including modification of open reading frames, production of alternatively spliced mRNAs, alterations of transcription factor binding sites and structural gains or losses within the regulatory regions. Consortium efforts such as the 1000 Genomes Project (1000GP) estimate a typical genome contains 2.1–2.5K SVs, affecting ~20M bases (~5–6 times that of SNPs in a typical genome). Beyond “simple” SVs, there is a growing appreciation of “complex” SVs in human genomes, which vary considerably in their architecture, ranging from small-scale insertions/deletions at larger SV breakpoints to complex patterns of rearrangements between distinct loci and/or different chromosomes[1](#_ENREF_1). Through 1000GP, we found that a large fraction of SV events, observed at fine resolution, have much higher breakpoint complexity than previously estimated—suggesting that complex SVs, like simple SVs, are widespread in human genomes.



**Figure 1.** **Overall research plan for the JAX CSVA.**

Through efforts such as the ongoing 1000GP, we are compiling vital whole-genome data that will form the basis for comprehensive analyses of human genetic variation and will address current gaps in understanding of complex diseases. In many disease contexts, known common single nucleotide variants (SNVs) account for only a small percentage of phenotype variability. Because SVs are common, larger in size and highly structurally diverse, they are poised 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 common diseases. Furthermore, their functional impact, especially in noncoding regions, has not been investigated systematically. 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) functionally interpreting their origins and phenotypic effects, and iii) establishing associations between specific SVs and disease.

We seek support to establish The Jackson Laboratory Center for Structural Variation Analysis (JAX CSVA), to advance the overarching goals of the GSP through computationally driven discovery, functional validation and characterization of disease-associated SVs (**Figure 1**). We will integrate novel and powerful tools for high-resolution SV discovery and, in collaboration with the primary data-producing centers of the GSP, 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 develop novel methodologies for functional annotation of variants and characterization of associated biological processes (Aim 2); these studies 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 200K samples of the GSP, which will provide the necessary statistical power for meaningful genotype-phenotype associations for disease-based SV association studies (Aim 3). This will allow us to make inferences about human population structure and adaptation at a scale much greater than anything attempted so far. Our deliverables will be the largest library of validated SVs discovered in humans, together with an unprecedented platform of pipelines for comprehensive, high-resolution and large-scale SV analysis.

**INNOVATION**

The originality of the JAX CSVA lies in the integration of cutting-edge computational methodologies—pioneered by the group—into a comprehensive platform for novel SV discovery, characterization and association with common human diseases. It is well known that our ability to generate large-scale genomic data is far outstripping our ability to analyze it at the scale and resolution required to make definitive functional associations. This issue is particularly relevant in the context of complex SVs, for which important details of their origin and functional effects cannot be appreciated without tools for analysis at nucleotide resolution. Furthermore, the present approach combines high-resolution SV analysis balanced against the scale required for adequately powered association analyses. Our proposed detection and genotyping strategy provides higher power and resolution for investigating association between SVs spanning a large size spectrum and various phenotypes, surpassing previous standard approaches employed in several major SV association studies. Briefly, the key innovations of our approach are: **1)** Development of a scalable pipeline incorporating the latest, cutting-edge SV detection and integration tools, with a focus on high-resolution classification of complex SVs. **2)** Tools for annotating variants with functional data from coding and non-coding parts of the genome. **3)** Tools for mechanistic interpretation of SVs across different classes, allowing us to make inferences about population structure and human adaptation and evolution. **4)** Association tests that integrate weighting methods for various biological considerations, such as allele frequency and impact score, to a generalized linear model for capturing subtle association signals often missed by conventional approaches. **5)** Genotyping the library of functionally and genetically relevant SVs across the entire cohort of GSP samples for well-powered genotype-phenotype associations in a disease context. **This systematic review 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.**

**CENTER STRUCTURE AND ADMINISTRATION**

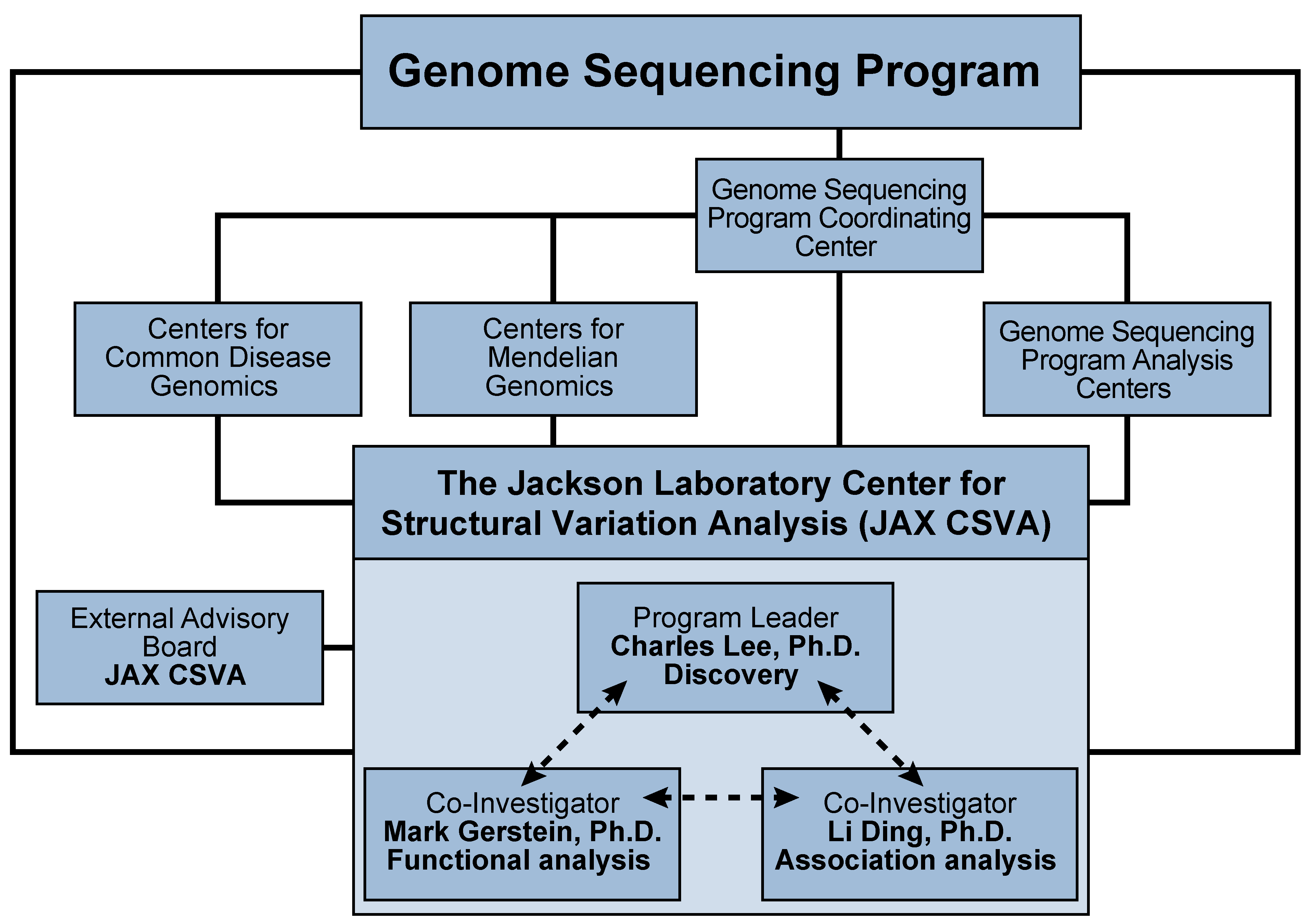


Figure 2. Structure of the JAX Center for Structural Variation Analysis

We envision the JAX CSVA as a nexus that will connect the data-producing and analysis centers of the GSP **(Figure 2).** By focusing on the discovery and analysis of SVs—a widespread form of genomic variation observed across common and Mendelian diseases—we expect to inform the analyses conducted by other GSPACs. We anticipate frequent communication and exchange of data sets, analytical tools, and results, as this will be critical for maximizing the utility of the data and extracting meaningful biological insights. These exchanges will also be critical for facilitating program-related objectives, i.e., towards the definition of new standards and controls for complex common disease studies, which similarly cut across the goals of individual GSP components.

Scientists participating in the JAX CSVA 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 (Ding). Each also brings significant experience in leading (1000GP SV group; Lee) and participating in (1000GP, Lee/Gerstein/Ding; ENCODE, Gerstein; ICGC, Gerstein/Ding) large-scale sequencing consortia. Notably, under Dr. Lee’s leadership, the 1000GP SV project identified SV events in 2,500 healthy genomes and helped define the methodology for identifying and characterizing SVs from whole genome sequencing datasets. This experience will be critical for facilitating cross-program goals, i.e., defining metrics for robust and comprehensive complex disease studies, and for ensuring seamless integration of the JAX CSVA within the broader GSP. The team includes investigators with broad knowledge and leadership experience in large-scale genomic sequencing consortia (Weinstock) and in mathematical and biostatistical tool development for the analysis of complex genomic datasets (Wendl, Chuang, Ye, Zhao). The group has the collective expertise to bring about the goals of the Center.

Finally, the proposed program will be supported by the extensive computational infrastructure at The Jackson Laboratory for Genomic Medicine, the site for the proposed JAX CSVA. Generous institutional commitments (see letter: Liu) towards development of the JAX Cloud will furnish the data storage and computational power needed for the formidable requirements of the project. The JAX CSVA will further benefit from the information technology, computational, bioinformatics, and software expertise resident at JAX and will act as a hub for all intellectual, computing and resource sharing aspects of the proposed program, both among its members and between centers of the broader GSP.

**RESEARCH STRATEGY**

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

***Rationale.*** Complex SV events are outside the design scope of available SV methods, yet are often of high impact and important for disease studies. We propose the Integrated Analysis of Structural Variants(*iASV*), a pipeline of tools developed by our group to discover SVs in thousands of sequenced genomes. It would be applied to the discovery cohort of individuals being sequenced by the Centers for Common Disease Genomics (CCDG) and Centers for Mendelian Genomics (CMG). Using breakpoint assembly methods (see below) we will perform *in silico* validation of the SV events and use the assembled contigs to investigate the inherent complexity prevalent at breakpoints. These studies will deliver the largest library of validated SVs discovered in humans and allow us to make novel biological inferences at the population level and in disease-specific contexts.



Figure 3. 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.

***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 and have developed a large toolbox of complementary tools and methods: ***(i) Read depth–based tool*** *(CNVNator).* We developed CNVnator[2](#_ENREF_2" \o "Abyzov, 2011 #587) 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 wide ranges 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*** *(BreakDancer, Meerkat, Hydra-Multi, Pindel).* Meerkat[3](#_ENREF_3" \o "Yang, 2013 #588), Hydra-Multi[4](#_ENREF_4) and BreakDancer[5](#_ENREF_5" \o "Fan, 2014 #60) 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[6](#_ENREF_6" \o "Ye, 2009 #16) utilizes a pattern-growth approach to detect large deletions and insertions from whole genome sequence (WGS). These methods have been successfully applied to hundreds of cancer genomes[3](#_ENREF_3),[7](#_ENREF_7). ***(iii) Split read alignment–based tools*** *(SRM, SRIC and BreakSeq).* We also developed SRM[8](#_ENREF_8) and SRIC[9](#_ENREF_9) for the high-resolution identification of SV events from WGS datasets. These tools provide base-pair resolution of breakpoints—a valuable feature that enables functional interpretation of the biology at play in these events.

We also developed the *Genome Variant Investigation Platform (GenomeVIP, manuscript in preparation)* for analyzing very large cohorts. It is a pipeline framework that manages the use of individual tools for germline, somatic, and *de novo* SV analysis in the cloud or on traditional high-performance compute clusters.

*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 described[7](#_ENREF_7), we used assembly-based methods like SGA[10](#_ENREF_10) or TIGRA-SV[11](#_ENREF_11) for generating sequence contigs at the breakpoint. Aligning these contigs back to the genome in the expected location and orientation validates the SV call (**Figure 3**). Using this method, we validated 64.8% of somatic breakpoints and 58.5% of germline control breakpoints[7](#_ENREF_7). We also developed AGE[12](#_ENREF_12), which performs sequence alignment at regions flanking SVs while considering large deletion and insertion blocks, which cannot be handled by conventional global and local 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[7](#_ENREF_7),[13](#_ENREF_13). We developed PEMer[14](#_ENREF_14" \o "Korbel, 2009 #603) for identifying complex rearrangements from WGS datasets.In another study, we comprehensively characterized complex SVs from a large cohort of TGCA WGS datasets[7](#_ENREF_7) 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 broad classes (**Figure 4a**). The most common class, *Ins with Dup and Del*, (N=501, 30%), exhibited a recognizable duplicated sequence interval within the respective inserted sequence. Not all sites fit neatly into the classes depicted in Figure 4a, 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 4b**), likely involving DNA replication errors[15](#_ENREF_15),[16](#_ENREF_16). These results reveal the extensive complexity of SV breakpoints and highlight the importance of mining this complexity at fine resolution for interpreting the biology of SVs.

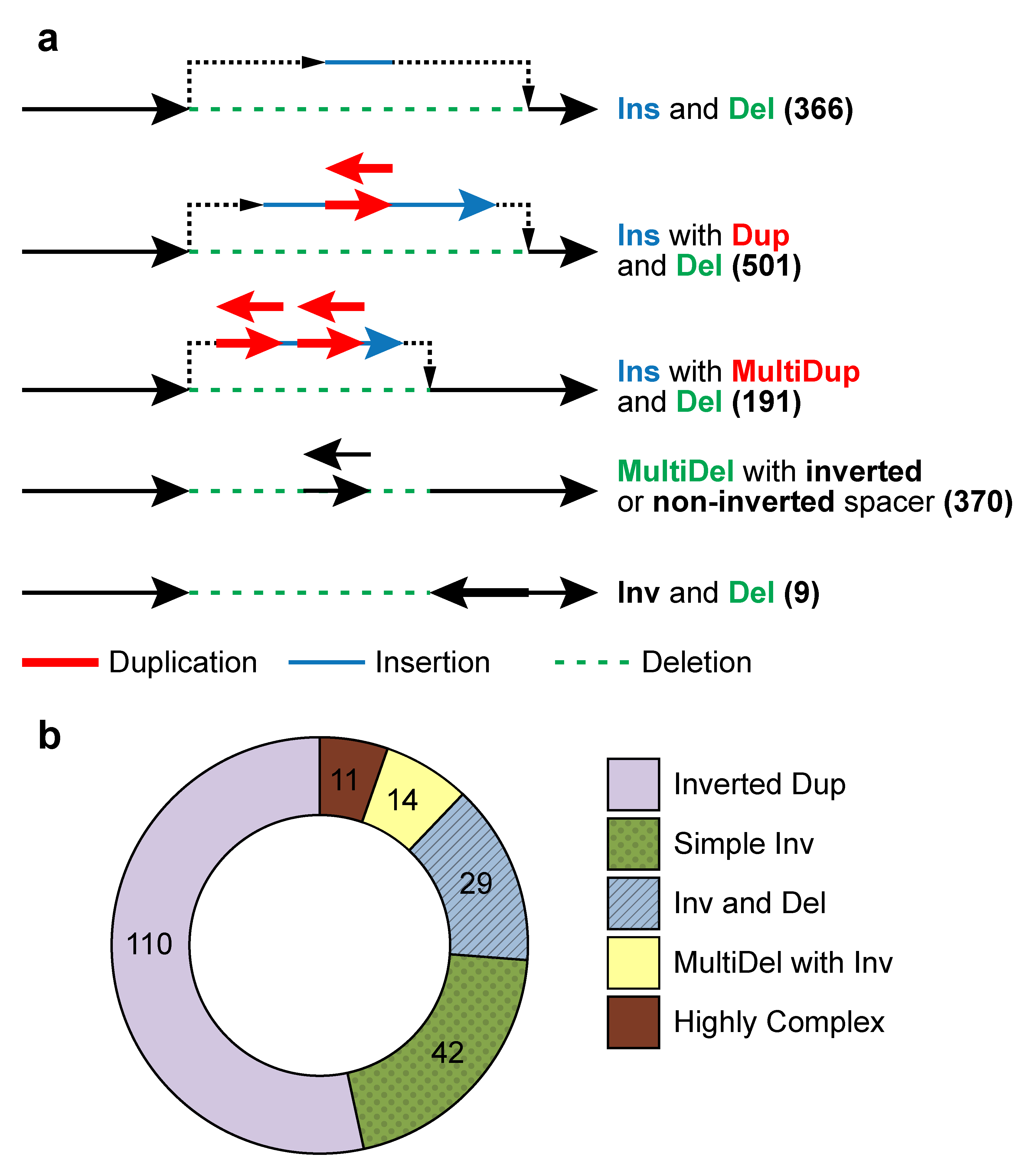


Figure 4. Structural variant complexity. a) We analyzed complexity of ~30000 deletions from the 1000 Genomes consortium phase 3 dataset, and characterized the events into several different categories based on the amount of complexity that was observed at the locus. b) Similar study of breakpoint complexity was performed for inversion events and showed much higher levels of complexity than expected.

***Research Plan.*** We plan to improve and integrate the described tools to identify and classify somatic and germline SVs across WGS datasets from the primary data-generating centers of the GSP. The new pipeline, *iASV*, will deliver i) integrated identification of a broad spectrum of SV types created by different molecular mechanisms; ii) compatibility with second- and third-generation-sequencing technologies; and iii) breakpoint resolution identification based on AGE (and other tools) and local assembly for *“in silico”* validation of the SV event.

*Sample selection.*Data storage and compute requirements preclude SV discovery on the whole GSP sequencing cohort. Based on our power calculations (Aim 3) we will select a discovery cohort of 10K individuals across multiple CCDGs and CMGs for *de novo* SV calling. 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). 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 with emphasis on genotype concordance for overlapping sites. The proposed iASV pipeline (**Figure 5**) for population-level 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, BreakDancer, Pindel, Hydra-Multi, Meerkat, SRM, SRIC and PEMer. 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 confirmation by multiple methods. (4) Unified methods for SV merging, genotyping and phasing using the lessons learnt from Phase 3 of the 1000GP[17](#_ENREF_17). (5) Validation for iASV and 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.

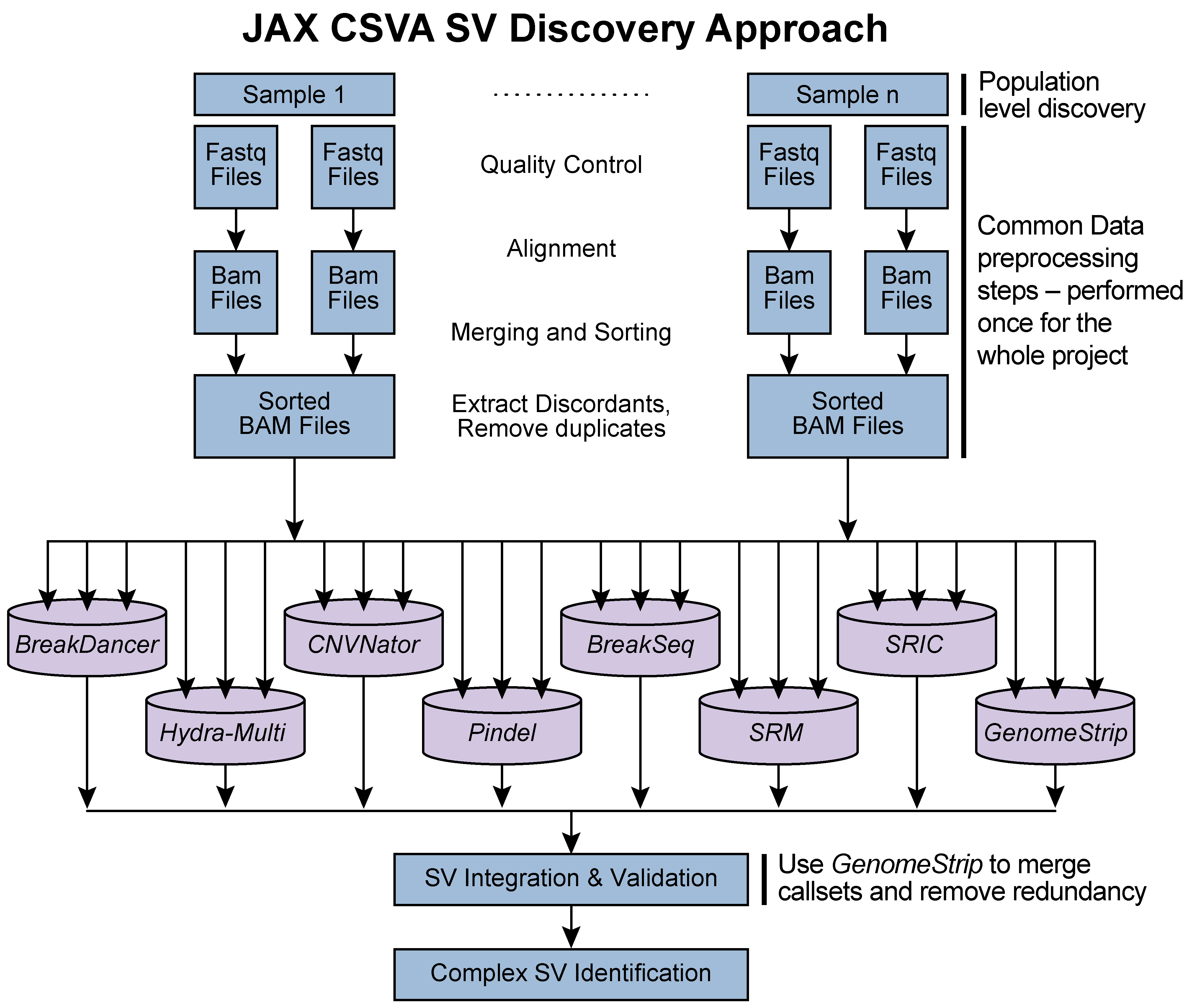


Figure 5. The integrated Analysis of Structural Variation (iASV) methodology.

The SV calling shall be performed in three phases:

*Phase 1—Calibration:* The pipeline will use a machine learning approach to calibrate and test the parameters of the different SV calling methods. We will initially focus on 100 “simulated truth” (ST) samples generated using WGSim (https://github.com/lh3/wgsim), 50 deep coverage “known truth” (KT) samples from 1000GP SV Project[17](#_ENREF_17) and 200 test cohort (TC) samples (from the TCGA consortium). These datasets all have some known true positive SVs and will be given different weights in the eventual determination of pipeline parameters depending on the level of confidence in the associated SV set (ST>KT>TC).

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

*Phase 3—Discovery:* The optimized system will be run on 10K of the proposed 200K individuals sequenced by the CMGs and CCDGs.

*Calibration of method using known sites.* Hundreds of sites across the human genome are polymorphic in a large fraction of the population[18](#_ENREF_18),[19](#_ENREF_19). Copy number variation can have diverse effects on phenotype[20-22](#_ENREF_20). For example, the opsin gene locus on chrX is highly polymorphic in ~11% of individuals[23](#_ENREF_23) and could explain the high incidence of male color blindness[24](#_ENREF_24). Phase 3 of 1000 GP SV project[17](#_ENREF_17) 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). We will integrate the same methodology into the iASV and process every discovered SV site for validation.

*Complex SV identification.*We will use two methods for complex SV identification. The first[7](#_ENREF_7) 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[17](#_ENREF_17) 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 SV calling methods.

*Generating a population-level reference set of SVs.*We expect that several different population groups will be represented in the complete cohort of individuals being sequenced at the CCDGs and CMGs. The resulting set of validated SVs from this aim (identified from the discovery cohort) will be further stratified according to underlying population substructure. The population-specific reference set of SVs will allow us to extend the observations from 1000GP Phase 3[17](#_ENREF_17) and will be critical for the population and disease-level association analyses proposed in Aim 3.

*Data access strategies: The JAX CSVA 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 CSVA Cloud, which will be available to all members of the JAX CSVA. 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 1700 cores and 1.4 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.* After initial method development and analysis, we plan to disseminate methods to the broader research community using the cloud paradigm. 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 we may use to assist with initial variant discovery and to download results to local disks for subsequent analyses.

JAX is now evaluating commercial genomics cloud service providers (CSPs) to partner with (see letters: Seven Bridges Genomics and IBM) and is now recruiting 1–2 full time employees for this effort. These activities are independent of this U01 proposal. JAX’s will choose a platform that will allow scientists without special training to analyze their datasets through a graphical interface for both local and cloud analysis methods. This goal parallels that of the U01, namely to ensure methods developed at the data center will be stably 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 GSP datasets. By making the iASV pipeline available as a community resource, we expect this work will propel future genome-level SV analyses.



Figure 6. Overview of the functional prioritization and annotation pipeline.

***Pitfalls and alternative approaches.*** A major challenge for this study is the lack of knowledge about the size and diversity of patient cohorts and details of common/mendelian diseases being studied as part of the GSP. We also do not know what kind of phenotype data is being collected and what orthogonal data (genomic or proteomic) might be available for these patient cohorts. In response, we will i) leverage the extensive experience of the team to handle complex datasets (see Prelim data section) and ii) design iASV to robustly handle diverse and complex datasets of the type that might be generated by the GSP.

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

***Rationale.*** Complex SVs are associated with genetic diseases and are responsible for more nucleotide variation in the human genome than SNPs. Despite their relevance, little is known about their functional impact at a genome-wide level. These events are disproportionately observed in the non-coding part of the genome; hence, comprehensive assessment of the functional impact of SVs will likely require the integration of large-scale data resources such as ENCODE, 1000GP, and GTEx. We anticipate that this proposal will catalogue the largest number of SVs so far; therefore, new methods to functionally prioritize SVs and select appropriate subsets for association studies will be necessary (**Figure 6**).

***Preliminary data.***

*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 have studied the distinct features of SVs originating from different mechanisms[25](#_ENREF_25),[26](#_ENREF_26). The most notable, non-allelic homologous recombination (NAHR), is associated with active enhancers and an open chromatin environment. Our analysis also showed that microinsertions flanking non-homologous breakpoints are templated from late replicating genome loci with characteristic distances from breakpoints. These results not only shed light on mechanisms underlying SV formation but also indicated differences in functional impacts of different SV types.

We further performed SV mechanism annotations for the 1000GP Phase 3 deletions using BreakSeq[27](#_ENREF_27" \o "Lam, 2010 #72), categorizing 29,774 deletions by their creation mechanisms. Among these, NHR proved to be the most prevalent mechanism (~73% of all categorized deletions)[17](#_ENREF_17).

*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 mutations according to synonymous, missense, nonsense or splice-site-disrupting changes[28](#_ENREF_28). We annotated variants from 1,092 humans within 1000 GP[25](#_ENREF_25) and observed that genes tolerant of loss-of-function (LoF) mutations are under the weakest selection and cancer-causal genes under the strongest. In phase 3 of the 1000GP we found that a typical genome contains ~150 LoF variants and also discovered a significant depletion of SVs (including deletions, duplications, inversions and multiallelic CNVs) in CDS, UTRs and introns of genes compared to a random background model, which implies strong purifying selection.

*Tools for evaluating functional impact of variation in non-coding (nc) RNAs and regulatory regions*. We have 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[29](#_ENREF_29" \o "Lu, 2011 #620). 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[30](#_ENREF_30).

To better understand non-coding (NC) regulatory regions, we developed tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. PeakSeq and MUSIC identify regions bound by TFs and chemically modified histones[31](#_ENREF_31),[32](#_ENREF_32). PeakSeq has been widely used in consortium projects such as ENCODE[31](#_ENREF_31),[33](#_ENREF_33). MUSIC 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[32](#_ENREF_32). 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[34-37](#_ENREF_34). In order 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[30](#_ENREF_30) [37](#_ENREF_37),[38](#_ENREF_38). We used metrics, such as diversity and fraction of rare variants, to characterize selection pressure on various classes and subclasses of functional annotations[30](#_ENREF_30). We have also defined variants that are disruptive to a TF-binding motif in a regulatory region[33](#_ENREF_33).

*Tools for helping annotate functional impact based on network and allelic expression analyses.* We found that functionally significant and highly conserved genes tend to be more central in various biological networks[39](#_ENREF_39) and are positioned at the top of regulatory networks[38](#_ENREF_38). Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery[39](#_ENREF_39),[40](#_ENREF_40)). Incorporating multiple network and evolutionary properties, we developed a NetSNP[39](#_ENREF_39" \o "Khurana, 2013 #637) 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 analyzing network hierarchy[38](#_ENREF_38). To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM[41](#_ENREF_41)). Finally, we also developed AlleleSeq[42](#_ENREF_42" \o "Rozowsky, 2011 #644) for the detection of candidate variants associated with allele-specific binding (ASB) and allele-specific expression (ASE). These tools are based on the construction of a personal diploid genome sequence (and corresponding personalized gene annotation) using genomic sequence variants and can be used to prioritize variations disrupting allelic activity.

*FunSeq: Tools for integrated functional prioritization.*We recently developed a prioritization pipeline called FunSeq[25](#_ENREF_25),[43](#_ENREF_43) that identifies annotations under strong selective pressure as determined using genomes from many individuals from diverse populations. FunSeq links each NC single-nucleotide mutation to target genes and prioritizes it based on scaled network connectivity. It 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 loss-of and gain-of function events). We further developed FunSeq (**FunSeq2**) and identified ~100 nc candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples[25](#_ENREF_25).

*Tools for identifying enrichment of variations in coding and non-coding regions.* We have worked on statistical methods for analysis of nc regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in nc elements by comparing observed mutation counts with expected counts under a whole genome background mutation model. LARVA also includes corrections for biases in mutation rate owing to DNA replication timing. For coding region analysis, we developed MuSiC[44](#_ENREF_44" \o "Dees, 2012 #648) to analyze genetic changes using standardized sequence-based inputs, along with multiple types of clinical data, to establish correlations among variants, affected genes and pathways, and to ultimately separate commonly abundant passenger events from truly significant events.

***Research plan.***We will extend the current FunSeq (**FunSeq2**) prototype to enable the identification of key SVs with high functional impact. Our new pipeline is called *SVIM* (Structural Variation IMpact). We will evaluate the impact score for each SV identified in Aim1 taking into account the functional annotation of the affected genomic region, as well as the fraction of functional elements (i.e. genes, ncRNAs and nc regulatory elements) overlapped by the SV. Furthermore, the impact score will also depend upon SV type (i.e. deletion, duplication, inversion & translocation).

For a given SV, which belongs to a particular SV type, we will evaluate 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 (Fsvtype, class). We will divide genomic elements into three categories (coding region, non-coding region, and TF binding site) and assign relative score to them (Scoding, Snon-coding, STFBS), which will vary for different SV types. Relative scores F and S, will be defined for class and functional elements in an analogous way to FunSeq2 tool[25](#_ENREF_25).

SVs will be assigned an impact score by taking the sum over the product between weights of overlapping classes and score of overlapping functional elements. The score (ISorig) will also be upweighted based on activity of the affected region. The upweight factor is comprised of the product of three factors (g) i.e. allelic activity, network connectivity and ubiquitous transcription. Significance level of am Impact score (ISorig) will be estimated by running a 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 LoF variants with mis-mapping, functional, evolutionary and network features. We will quantify the confidence of 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. Finally, we will develop a machine-learning method to quantify whether LoFs will cause benign, recessive or dominant disease-causing effects. Given that most rare variants are heterozygous, developing methods to differentiate benign rare variants from disease-causing variants in terms of those that can lead to recessive or dominant disease are much needed.

*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 are sensitive to mutation in the human population, 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[45-49](#_ENREF_45). 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 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 noncoding RNA.

*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 creation of new motifs or whether they lead to the breakage of an already existing motif. 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 first update the TF binding nc elements using better enhancer definition provided by the Epigenome Roadmap[50-52](#_ENREF_50), and ENCODE. Furthermore, we will develop a new machine-learning framework that utilizes pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA) to predict active enhancers across different tissues.

*Further variant prioritization based on networks, tissue specificity, and allelic activity*. After performing annotation-based assessment of identified SVs, the following functional features will be used for their further 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 specificity.* We will evaluate the impact of SVs in the epigenetic context in order 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.* Personal allelic variants (rare and common) will be aggregated into a reference set of genomic elements displaying allele-specific behavior and each element will be assigned an “allelicity” score based on enrichment of allelic variants both within the element and across individuals (with allelic variants in a consistent allelic direction). We will then develop a prioritization scheme for SVs overlapping these allelic elements.

***Expected results.***We propose to develop a software solution, called **SVIM**, to estimate the impact score of the SV set produced in Aim 1. We will prioritize these SVs and forward them to Aim 3 (genotype and association) to further classify population impact of these variants. We plan to make the prioritization results broadly available; therefore, SVIM will incorporate the impact score into a standard Variant Call Format (VCF).

***Pitfalls and alternative approaches.*** We anticipate that the greatest pitfall is (i) possibly overwhelming number of SV to be discovered in Aim 1 and (ii) the data that will be pre-processed to generate reliable annotation of nc component of analysis. In order to overcome (i), we plan to gradually process the results dividing analyzes into SV type tiers, initially prioritizing specifics type of SVs. SVIM will also be based on the data context to efficiently prioritize variants 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 integrate large-scale publicly available data resources, such as SVs from 1000 Genomes project[53](#_ENREF_53), conservation data from Bejerano *et al.*[54](#_ENREF_54) and Cooper *et al.*[55](#_ENREF_55), functional genomics data from ENCODE[33](#_ENREF_33) and Roadmap Epigenomics Mapping Consortium[56](#_ENREF_56).

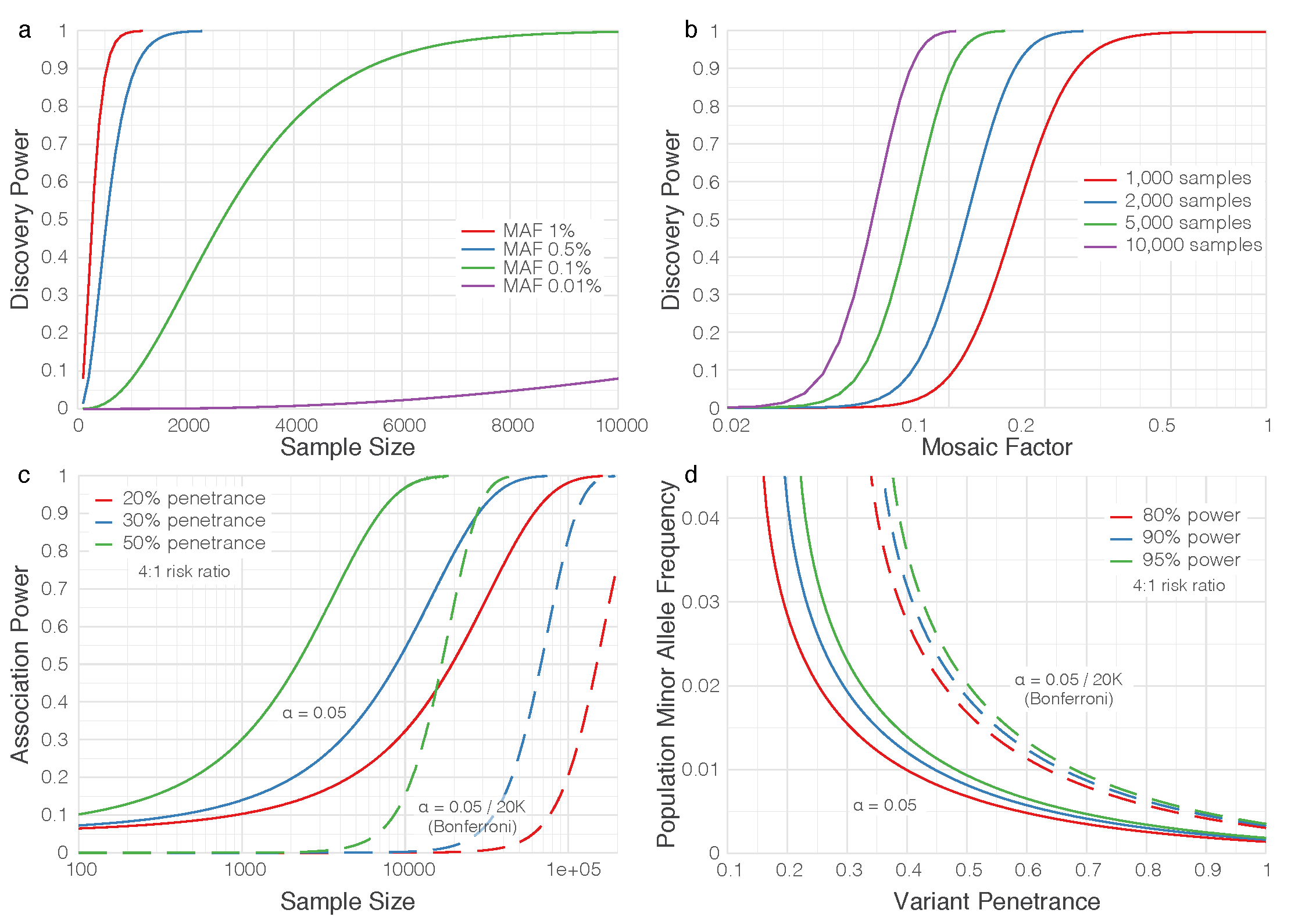
**Aim 3. Scaling up to 200K samples and associating SVs with common and rare diseases.**

***Rationale.*** As many of the high-impact structural variants will be relatively rare that conventional association tools cannot readily, robustly handle, we will not only discover important SVs, but will develop a new association pipeline tailored 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 to this undertaking.

**Preliminary Results**

*Power analysis for sample selection and association.* An important aspect will be selecting of a subset of the 200K samples projected to be sequenced for full SV analysis. This “discovery phase” 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[57](#_ENREF_57) to estimate the discovery sample size. Briefly, 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.

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).

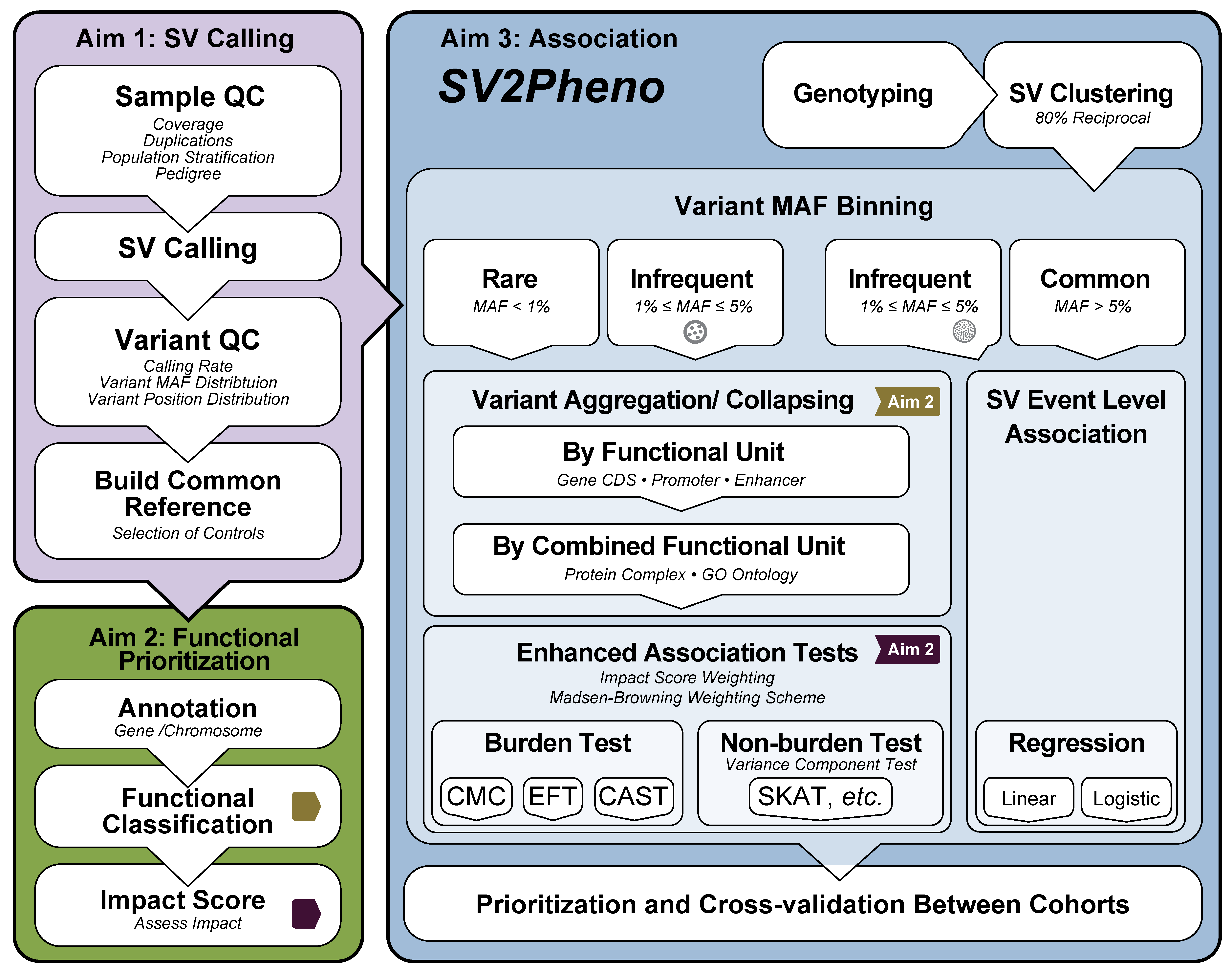


Anticipated parameters for the Illumina data to be generated for this project are 30X coverage per genome, average insert size of 400bp-600bp (20% coefficient of variation), 150bp reads, event detection based on ≥3 split reads or ≥5 discordant read pairs, and observation in at least 3 samples to constitute “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.

The second aspect of “power” is variant-disease association. The issues are well-known[58](#_ENREF_58), enabling the following “baseline” estimates of association power. General consensus[58](#_ENREF_58) 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 200K 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 what solid associations can be established for.

*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 sports popular aggregation tests, including burden tests such as the Combined Multivariate Collapsing (CMC)[58](#_ENREF_58), Exclusive Frequency Test (EFT)[59](#_ENREF_59), Total Frequency Test (TFT)[59](#_ENREF_59), and Cohort Allele Sum Test (CAST)[60](#_ENREF_60), and variant component tests such as the Sequence Kernel Association Test (SKAT)[61](#_ENREF_61). 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.

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.



**Research Plan.** SVs are characterized by size, type, penetrance, and multiple alleles. We plan to genotype all SVs detected in 10K discovery samples (**Aim 1**) across all ~200K 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[27](#_ENREF_27" \o "Lam, 2010 #72)** 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](#_ENREF_70" \o "Handsaker, 2011 #73)** 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 an expected data footprint of ~50PB 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](#_ENREF_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 2 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](#_ENREF_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. (2) Variants are known to be associated with various diseases[73-75](#_ENREF_73), 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 , 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 4 considerations: the Madsen-Browning equation[76](#_ENREF_76) to account for allele frequency, consideration of “direction” (negative association) using e.g. aspects of the Pan-Shen approach[77](#_ENREF_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, 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[78](#_ENREF_78).

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 GSP 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[11](#_ENREF_11). 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 (**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 potentially 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 any 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[61](#_ENREF_61),[76](#_ENREF_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 JAX CSVA **SV2Pheno** association pipeline and its associated/support 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 TopMedcenters. 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.

**CONTRIBUTIONS TO CROSS-PROGRAM GOALS**

The analyses to be undertaken by the JAX CSVA will contribute to the two primary cross-program goals of the GSP. These methodologies and analysis tools are integral to the investigator-driven activities of the proposal; thus, it will not be necessary to prioritize them as separate initiatives. This strategy insures that these program-related goals will be achieved in line with Center-specific goals.

*Delineating comprehensiveness in common disease studies.* The JAX iASV approach will allow us to integrate samples from across the various centers of the GSP into a single meta-analysis of SV across thousands of genomes. This allows for biological interpretation across the width of the GSP and will enable investigators to answer questions about population structure and their impact on phenotype. Thus, the JAX CSVA will contribute to cross-program objectives by integrating data from across centers in a disease agnostic manner. Furthermore extensive calibration and optimization of the various tools that are part of the iASV, as well as the tight integration with cloud-based computing, will also help define the methodology and metrics for comprehensive study of SVs in future large-scale consortium efforts.

*Providing specifications for common controls.* JAX CSVA will combine our well-curated, genotyped SVs with calls from CCDG and CMG centers to perform a SV saturation analysis[79](#_ENREF_79) to assess the completeness of SV census across populations and disease types. We shall also integrate SVs with SNVs/indels generated by other centers as part of this proposal and we anticipate this effort will yield a larger set of association hypotheses outside the scope of our proposal, but perhaps well-suited for other proposals in the GSP program. Furthermore, we will share our **SV2Pheno** association pipeline and its embedded tools through cloud or local installation, which may help other projects in the GSP program. As described above, we shall choose appropriate controls for our SV discovery analysis based on population structure, and other confounding factors. These set of controls will be shared across the GSP and help other centers when considering the choice of controls for their analysis.

REFRENCES.

1 Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet* **28**, 43-53, doi:10.1016/j.tig.2011.10.002 (2012).

2 Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* **21**, 974-984, doi:10.1101/gr.114876.110 (2011).

3 Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-929, doi:10.1016/j.cell.2013.04.010 (2013).

4 Lindberg, M. R., Hall, I. M. & Quinlan, A. R. Population-based structural variation discovery with Hydra-Multi. *Bioinformatics* **31**, 1286-1289, doi:10.1093/bioinformatics/btu771 (2015).

5 Fan, X., Abbott, T. E., Larson, D. & Chen, K. BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **2014**, doi:10.1002/0471250953.bi1506s45 (2014).

6 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

7 Malhotra, A. *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome research* **23**, 762-776, doi:10.1101/gr.143677.112 (2013).

8 Malhotra, A. *et al.* Ploidy-Seq: inferring mutational chronology by sequencing polyploid tumor subpopulations. *Genome Med* **7**, 6, doi:10.1186/s13073-015-0127-5 (2015).

9 Zhang, Z. D. *et al.* Identification of genomic indels and structural variations using split reads. *BMC genomics* **12**, 375, doi:10.1186/1471-2164-12-375 (2011).

10 Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome research* **22**, 549-556, doi:10.1101/gr.126953.111 (2012).

11 Ryslik, G. A. *et al.* A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. *BMC bioinformatics* **15**, 231, doi:10.1186/1471-2105-15-231 (2014).

12 Abyzov, A. & Gerstein, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* **27**, 595-603, doi:10.1093/bioinformatics/btq713 (2011).

13 Weckselblatt, B. & Rudd, M. K. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet*, doi:10.1016/j.tig.2015.05.010 (2015).

14 Korbel, J. O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* **10**, R23, doi:10.1186/gb-2009-10-2-r23 (2009).

15 Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**, 551-564, doi:10.1038/nrg2593 (2009).

16 Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS genetics* **5**, e1000327, doi:10.1371/journal.pgen.1000327 (2009).

17 Sudmant, P. H. An integrated map of structural variation in 2,504 human genomes. *Nature* **Accepted, in print** (2015).

18 Wong, K. K. *et al.* A comprehensive analysis of common copy-number variations in the human genome. *American journal of human genetics* **80**, 91-104, doi:10.1086/510560 (2007).

19 Bailey, J. A., Kidd, J. M. & Eichler, E. E. Human copy number polymorphic genes. *Cytogenet Genome Res* **123**, 234-243, doi:10.1159/000184713 (2008).

20 Iskow, R. C., Gokcumen, O. & Lee, C. Exploring the role of copy number variants in human adaptation. *Trends Genet* **28**, 245-257, doi:10.1016/j.tig.2012.03.002 (2012).

21 Hehir-Kwa, J. Y., Pfundt, R., Veltman, J. A. & de Leeuw, N. Pathogenic or not? Assessing the clinical relevance of copy number variants. *Clin Genet* **84**, 415-421, doi:10.1111/cge.12242 (2013).

22 Almal, S. H. & Padh, H. Implications of gene copy-number variation in health and diseases. *J Hum Genet* **57**, 6-13, doi:10.1038/jhg.2011.108 (2012).

23 Drummond-Borg, M., Deeb, S. S. & Motulsky, A. G. Molecular patterns of X chromosome-linked color vision genes among 134 men of European ancestry. *Proc Natl Acad Sci U S A* **86**, 983-987 (1989).

24 Vollrath, D., Nathans, J. & Davis, R. W. Tandem array of human visual pigment genes at Xq28. *Science* **240**, 1669-1672 (1988).

25 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).

26 Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* **6**, 7256, doi:10.1038/ncomms8256 (2015).

27 Lam, H. Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature biotechnology* **28**, 47-55, doi:10.1038/nbt.1600 (2010).

28 Habegger, L. *et al.* VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28**, 2267-2269, doi:10.1093/bioinformatics/bts368 (2012).

29 Lu, Z. J. *et al.* Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome research* **21**, 276-285, doi:10.1101/gr.110189.110 (2011).

30 Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. & Gerstein, M. B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* **39**, 7058-7076, doi:10.1093/nar/gkr342 (2011).

31 Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology* **27**, 66-75, doi:10.1038/nbt.1518 (2009).

32 Harmanci, A., Rozowsky, J. & Gerstein, M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol* **15**, 474, doi:10.1186/s13059-014-0474-3 (2014).

33 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

34 Cheng, C. *et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* **12**, R15, doi:10.1186/gb-2011-12-2-r15 (2011).

35 Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research* **22**, 1658-1667, doi:10.1101/gr.136838.111 (2012).

36 Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448, doi:10.1038/nature13424 (2014).

37 Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48, doi:10.1186/gb-2012-13-9-r48 (2012).

38 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).

39 Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886, doi:10.1371/journal.pcbi.1002886 (2013).

40 Kim, P. M., Korbel, J. O. & Gerstein, M. B. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**, 20274-20279, doi:10.1073/pnas.0710183104 (2007).

41 Cheng, C. *et al.* An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. *Genome Biol* **16**, 63, doi:10.1186/s13059-015-0624-2 (2015).

42 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522, doi:10.1038/msb.2011.54 (2011).

43 Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, doi:10.1186/s13059-014-0480-5 (2014).

44 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome research* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).

45 Blin, K. *et al.* DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **43**, D160-167, doi:10.1093/nar/gku1180 (2015).

46 Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* **42**, D92-97, doi:10.1093/nar/gkt1248 (2014).

47 Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141, doi:10.1016/j.cell.2010.03.009 (2010).

48 Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**, 654-665, doi:10.1016/j.cell.2013.03.043 (2013).

49 Garcia, D. M. *et al.* Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol* **18**, 1139-1146, doi:10.1038/nsmb.2115 (2011).

50 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

51 Ziller, M. J. *et al.* Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* **518**, 355-359, doi:10.1038/nature13990 (2015).

52 Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350-354, doi:10.1038/nature14217 (2015).

53 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

54 Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321-1325, doi:10.1126/science.1098119 (2004).

55 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**, 901-913, doi:10.1101/gr.3577405 (2005).

56 Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28**, 1045-1048, doi:10.1038/nbt1010-1045 (2010).

57 Wendl, M. C. & Wilson, R. K. Statistical aspects of discerning indel-type structural variation via DNA sequence alignment. *BMC genomics* **10**, 359, doi:10.1186/1471-2164-10-359 (2009).

58 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics* **83**, 311-321, doi:10.1016/j.ajhg.2008.06.024 (2008).

59 Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-872, doi:10.1126/science.1099870 (2004).

60 Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation research* **615**, 28-56, doi:10.1016/j.mrfmmm.2006.09.003 (2007).

61 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* **89**, 82-93, doi:10.1016/j.ajhg.2011.05.029 (2011).

62 Ferguson, J. *et al.* Statistical tests for detecting associations with groups of genetic variants: generalization, evaluation, and implementation. *European journal of human genetics : EJHG* **21**, 680-686, doi:10.1038/ejhg.2012.220 (2013).

63 Ballard, D. H., Cho, J. & Zhao, H. Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genetic epidemiology* **34**, 201-212, doi:10.1002/gepi.20448 (2010).

64 Chun, H., Ballard, D. H., Cho, J. & Zhao, H. Identification of association between disease and multiple markers via sparse partial least-squares regression. *Genetic epidemiology* **35**, 479-486, doi:10.1002/gepi.20596 (2011).

65 Chen, M., Cho, J. & Zhao, H. Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS genetics* **7**, e1001353, doi:10.1371/journal.pgen.1001353 (2011).

66 Hou, L., Chen, M., Zhang, C. K., Cho, J. & Zhao, H. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Human molecular genetics* **23**, 2780-2790, doi:10.1093/hmg/ddt668 (2014).

67 Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature genetics* **40**, 592-599, doi:10.1038/ng.118 (2008).

68 Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124, doi:10.1038/nature11582 (2012).

69 Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223, doi:10.1038/nature12141 (2013).

70 Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics* **43**, 269-276, doi:10.1038/ng.768 (2011).

71 Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034, doi:10.1093/bioinformatics/btv098 (2015).

72 Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics* **91**, 224-237, doi:10.1016/j.ajhg.2012.06.007 (2012).

73 Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449, doi:10.1126/science.1138659 (2007).

74 Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-543, doi:10.1126/science.1155174 (2008).

75 Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nature genetics* **43**, 838-846, doi:10.1038/ng.909 (2011).

76 Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics* **5**, e1000384, doi:10.1371/journal.pgen.1000384 (2009).

77 Pan, W. & Shen, X. Adaptive tests for association analysis of rare variants. *Genetic epidemiology* **35**, 381-388, doi:10.1002/gepi.20586 (2011).

78 Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *American journal of human genetics* **95**, 5-23, doi:10.1016/j.ajhg.2014.06.009 (2014).

79 Schnabel, Z. E. The Estimation of Total Fish Population of a Lake. *American Mathematical Monthly* **45**, 348-352 (1938).