**SPECIFIC AIMS**

Genomic structural variations (SVs), such as deletions, duplications, insertions, inversions and translocations, comprise a powerful class of phenotype-shaping genetic variation. Significant advances in our ability to discover SVs in genomes, owing to the advent of next-generation sequencing technologies and large-scale sequencing efforts such as the 1000 Genomes SV Project (1000 GP), have enabled us to link specific SVs with many of humanity’s most challenging diseases, particularly cancer. However, these associations remain mostly correlative and do not provide mechanistic insight into the role SVs play in various diseases. Adding to the complexity are recent observations of complex SV events such as Chromothripsis and Chromoplexy that can cause large-scale disruptions to DNA structure. This is supported by new observations within the 1000 GP showing that a large fraction of SV events have much higher breakpoint complexity that previously estimated. These findings highlight the challenges of investigating the genomic mechanisms of disease associated with SVs, as most studies do not adequately cover the full spectrum of complex SVs despite their importance for disease biology.

The present proposal seeks to address this gap through the development of The Jackson Laboratory Center for Structural Variation Analysis (JAX CSVA) with collaboration between the leaders in the field of structural genetic variation. The JAX CSVA will be headed by Dr. Charles Lee as Center Director and will have major contributions by the research groups of Dr. Mark Gerstein and Dr. Li Ding (as senior P.I.). The other important contributors shall be Drs. George Weinstock, Jeff Chuang, [[MG-comment]],,hz Hong Yu, Ankit Malhotra and Zhengqing Ouyang. This effort grows out of a long-term association between the Lee, Ding & Gerstein labs. Working together within the framework of the 1000 Genomes SV project, the labs have worked together as a tightly integrated team towards the common goal of identifying the sources of genetic variation in human genomes. The team has an extensive and demonstrable track record in developing new algorithms for detecting and characterizing structural genomic variants (XXX CITE).

The JAX CSVA will establish a novel pipeline for discovering, validating and genotyping complex SV events from thousands of genomes that we will use to discover complex SVs from a cohort of genomes being sequenced at the Centers for Common Disease Genomics (CCDG) and Centers for Mendelian Genomics (CMG). We will focus our efforts on the space of sequence variants not covered by these large sequencing centers – especially [[MG-comment]],,large indels and SVs. We will then genotype this selection of complex SVs and indels in the full cohort of individuals sequenced as part of these efforts. Using novel methodologies we will functionally annotate the variants and perform genome-wide association studies (GWAS) in a disease-specific context. The investigator-led component of the Center will focus on identifying SVs and their functional impact on disease biology, while the collaborative component will seek to define the processes and metrics for future SV studies and publish a reference set of complex SVs and indels as a community resource. This novel, systematic and comprehensive approach to the investigation of complex SVs will yield valuable tools for future investigations, including a reference catalogue of germline events from thousands of individuals and a standard set of tools, pipelines and metrics for performing such studies in the future. In terms of novel biology, the JAX CSVA will highlight the prevalence and importance of complex SVs in disease and provide mechanistic and functional insight into complex SV formation and its effects. Towards these goals, we propose the following Specific Aims:[[MG-comment]]missing GWAS

**Aim 1. Build an integrative pipeline for large-scale discovery of complex structural variation.** Complex SV events are often overlooked by structural variation studies. We propose to build an integrated, efficient and scalable pipeline of tools developed by our group to discover all classes of structural variations in thousands of genomes. The pipeline would be applied to a selected cohort of individuals being sequenced as part of the Genome Sequencing Program (GSP). Using breakpoint assembly methods we will perform *in silico* validation of the SV events and use the assembled contigs to gain a handle on 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.

**Aim 2. Develop tools to analyze the functional impact of SVs.** We anticipate that most of SVs discovered in the human genome will not impact coding regions; thus, methods to evaluate the functional impact of SVs need to be genome-wide, including non-coding regions. We propose to develop a framework to evaluate SVs over three contexts: (1) Impacting protein coding genes; (2) Impacting non-coding RNAs; (3) Impacting non-coding regulatory regions such as Transcription Factor Binding Sites (TFBS). The impact score will take into account the varied ways that a SV can affect genomic elements (e.g. partial overlap or engulf) and will integrate conservation information, existing genomic annotations, and epigenetic and transcriptomic datasets from sources such as ENCODE, 1000 Genomes, and GTEx. Furthermore, we will upweight the impact score of SVs overlapping elements with ubiquitous activity, high network connectivity (i.e. hubs) and strong allelic activity (i.e. demonstrated functional sensitivity to variants).

**Aim 3. Scaling up to 200K samples and associating SVs with common and rare diseases.** [[MG-comment]]enlargeWe 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 framework to integrate the regulatory signals coming from SV regions and assess their potential to disrupt normal biological processes. We will also perform cis and trans GWAS between gene expression and SV events. 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.

**B. 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 organismal phenotype due to the range of possible consequences that include modification of open reading frames, production of alternatively spliced mRNAs, alterations of transcription factor binding sites and structural gains or losses within the regulatory genomic space. Through consortium efforts such as the 1000 Genomes Project (1000GP), we are gaining increasingly comprehensive insight into the extent to which SVs are present across healthy populations. For example, as part of the 1000GP SV group, we estimate a typical genome contains 2.1–2.5K SVs, affecting ~20M bases of sequence (~5–6 times that of SNPs in a typical genome). This diversity in genomic architecture and functional effects underscores a central challenge in the field—to associate specific SVs with disease, particularly for those complex diseases for which common gene variants alone cannot account for observed phenotype variability.

Adding to the challenge is the growing appreciation of “complex” SVs in human genomes, involving various combinations of insertions, inversions, tandem duplications, translocations and deletions that evade simple classification. SVs are known to be associated with various diseases[7-9](#_ENREF_7), but their functional impact, especially in noncoding regions, has not been investigated systematically. Complex SVs 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 (ref). Extreme examples include the recently discovered phenomena chromothripsis and chromoplexy (XXX CITE), which can cause large-scale disruptions to DNA structure that are particularly appreciated in the context of cancer. [[MG-comment]]no somatic 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 are indeed widespread in human genomes. However, beyond extreme examples, the field remains in the dark in terms of how complex SVs arise in human genomes, the range of configurations they can take and, most importantly, how they give rise to relevant disease phenotypes.

Through large-scale consortium efforts, such as the ongoing 1000GP and soon-to-be established centers of the Genome Sequencing Program (GSP), we are gaining ever more access to vital whole-genome and exome data that will form the basis for comprehensive analyses of common (and complex) human genetic variation. Investigating SVs, and particularly complex SVs, could hold the key to a deeper, more mechanistic understanding of common diseases. However, it remains that most associations between SVs and disease are correlative at best and that, given that most studies do not capture the spectrum of complex SVs present in genomes, that this complexity is not adequately accounted for in association studies. A major contributor to this issue is …..

Mark/Li, please comment here.

Can you help develop this paragraph as the “knowledge/innovation gap” paragraph? Can you help to add text that addresses the following points:

* How our ability to create the data is far outpacing our ability to analyze it?
* How we do not have the tools to analyze the data in a way that allows us to draw correlative insights?
* How most analyses do not adequately account for complexity and/or lack the resolution for analyzing complex SVs?
* What analytical tools, computing infrastructure, etc are needed for adequately powered studies of complex SVs, association with disease phenotypes?
* What we as a group have done to address these issues already (very general, as I assume the details will be in the prelim data section), particularly in the context of the strengths of the group (scale, resolution, experience with large-scale consortia)? I think we should highlight our strengths here.

With this application, we seek support to establish The Jackson Laboratory Center for Structural Variation Analysis (JAX CSVA) for discovering, validating and genotyping complex SV events from a large cohort of genomes. We aim to use a novel approach to comprehensively discover all types of structure variants including complex SVs from a large number of genomes being sequenced at the Centers of Common Disease Genomics (CCDG) and Centers of Mendelian Genomics (CMG) (Specific Aim 1). Using novel methodologies such as SVIM etc we aim to functionally annotate the variants and reveal the biological processes that gave rise to them (Specific Aim 2). We also aim to develop SV2Pheno pipeline to genotype a subset of the relevant SV and Indels in the complete cohort of individuals and perform disease based association studies as well as make inferences about human population structure and adaptation at a scale much greater than anything attempted so far (Specific Aim 3)

The scientists assembled as part of the JAX CSVA are the leaders in the field of SV detection, functional interpretation and large scale data analysis. As a team they present a very unique blend of experience and expertise that will be unmatched and provide the GSP effort with novel biological insight as well as define the metrics for robust and comprehensive cross program analysis studies. This expertise is also highlighted by the huge success of the 1000 Genomes structural variation project - led by Dr. Lee and with important contributions from the Gerstein and Ding labs. The 1000 genomes SV project has, over a period of 5 years, not only comprehensively identified the full repertoire of SV events in 2500 healthy genomes [XXX cite phase 3 paper] but has also defined the methodology for identifying structural variations from whole genome sequencing datasets, characterizing them for function and gain insight into population structure and evolution. Further these tools in combination provide us with a comprehensive toolbox to identify SV at all scales and classes. XXX – Integrate with the other centers

**C. INNOVATION – I think this might be where I spend my time tomorrow; will try crafting this into better prose. But the innovation section goes here.**

* The proposed program takes advantage of the team’s unparalleled expertise in SV identification and analysis and combines it with powerful computing infrastructure at JAX-GM.
* The team offers particularly unique strengths in terms of its ability to assess SV complexity at scale required for the types of data to be generated through the GSP and the resolution necessary for analysis of complex SVs
* It is unlikely that any other group has the expertise for complex SV analysis at the scale and resolution we propose to achieve here…
  1. Integrated, comprehensive and scalable pipeline for SV calling with a focus on complex SV
  2. Make SV calls from a LARGE cohort of individuals (*statistically significant from power analysis*)
  3. Database of a reference set of validated (*in-silico*) SV (*with a focus on complex SV*)

2.1 Annotating variants with functional data from coding/non-coding parts of the genome

2.2 Mechanistic interpretation of SV across different classes

2.3 Make inferences about population structure and human adaptation / evolution

* 1. Genotyping the library of functionally and genetically relevant SV in the whole set of individuals
  2. GWAS analysis with SV and other genetic variants in a disease context

**Innovation:** Novel systematic review of complex SVs in many genomes. Reference catalogue of germline complex SV from many individuals. Mechanistic and functional insight into complex SV formation and its effects.

1. Detection pipeline with a focus on complex variations-- describe how it is unique and what it offers over and above other available approaches
2. Characterization of complex SV
3. Methods for associating complex structure variants with phenotypes and population structure [emphasis more in text]
4. Building a population reference for personal genome analysis
5. A large of number of novel functional SVs in non-coding sequences - how many novel SV’s do we expect to find with increasing number of individuals being sequenced.
6. Future direction - cross species comparisons of SV landscape !!!

****

**STRATEGY**

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

***Rationale.*** Complex SV events are often not adequately covered by the available automated SV calling pipelines, yet they are often of high impact and important for disease studies. We propose to build an integrated, efficient and scalable pipeline of tools developed by our group to discover SVs in thousands of sequenced genomes. The pipeline would be applied to a selected cohort of individuals (10–20K based on power analysis) being sequenced as part of the Centers for Common Disease Genomics (CCDG) and Centers for Mendelian Genomics (CMG). Using breakpoint assembly methods (described below) we will perform *in silico* validation of the SV events and use the assembled contigs to gain a handle on 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.

***Preliminary data.***

*A toolbox of methods for structural variation discovery.*The team of researchers assembled as part of the JAX CSVA are leaders in the field of SV detection and functional interpretation. As part of the 1000GP SV project we have provided the research community with an unprecedented and comprehensive set of germline SVs from more than 2,500 normal human genomes and have developed a large toolbox of complementary tools and methods. These tools can be broadly separated into three categories based upon the evidence used for calling SV events:

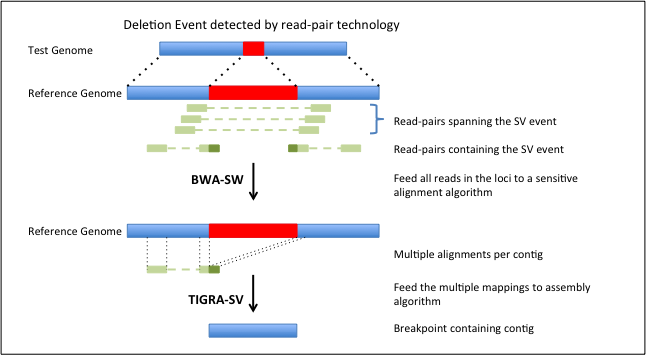
*i) Read depth­–based tools (i.e., CNVNator).* We developed CNVnator [21324876] for copy number variant (CNV) discovery and genotyping from individual and trio-sequencing datasets. CNVnator utilizes a read depth–based, mean-shift approach along with GC correction and novel multiple bandwidth partitioning to identify wide ranges of CNV events. We calibrated CNVnator by leveraging extensive validation exercises performed *via* the 1000GP. CNVnator can detect CNVs and provide genotype information on a population level, and can also accurately detect atypical CNVs including *de novo* and multi-allelic events.

*ii) Paired end–based tools (i.e., BreakDancer, Meerkat, Hydra-Multi, PEMer, Pindel).* Meerkat [23663786], Hydra-Multi [25527832] and BreakDancer [25152801] rely on clustering abnormally mapped paired-end reads to identify genomic loci with a signature for an SV event. Meerkat remaps soft clipped and unmapped reads to generate clusters that identify the breakpoints. Pindel [19561018] utilizes a pattern-growth approach to detect large deletions and insertions from WGS datasets. These methods have been successfully applied to hundreds of cancer genomes [23663786, 23410887]. As an example, we used Hydra-Multi to discover SVs from a large cohort of 64 whole genome sequencing (WGS) cancer datasets (and matching normals) from The Cancer Gene Atlas consortium (TGCA). We observed that a large proportion of identified SVs (1542/6179, ~25%) clustered into 154 complex SV clusters. These clusters were identified in 48/64 genomes (75%) representing all seven cancer types analyzed and were relatively evenly distributed across tumor types, with most tumors showing 1–5 clusters.

*iii) Split read alignment–based tools (i.e., SRM, SRIC and BreakSeq).* We also developed SRM [25729435] and SRIC [21787423] 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 e.g., the mutational mechanisms that give rise to these events.

For running an ensemble of SV discovery programs on a large scale, we also developed the Genome Variant Investigation Platform (GenomeVIP - CITE). GenomeVIP is a lightweight, web-driven, extensible genomics pipeline that enables germline, somatic, and *de novo* SV analysis to be performed in the cloud (in this case, Amazon’s cloud) or on traditional high-performance compute clusters. It deploys multiple computationally intense discovery tools capable of analyzing WGS and whole exome sequence (WES) data using best practices and customizable parameter sets.

*Breakpoint assembly tools for in silico validation.* Our group has also developed algorithms for identifying breakpoints at nucleotide resolution, thereby allowing us to validate SV breakpoints *“in silico”*. As described [23410887], we used assembly-based methods like SGA [XXX-Cite] or TIGRA-SV [XXX-Cite] for generating sequence contigs at the breakpoint. Aligning these contigs back to the genome in the expected location and orientation validates the SV call. Using this method, we validated 64.8% of somatic breakpoints and 58.5% of germline control breakpoints [23410887], with a median contig length of 862 bp. We have also developed AGE [21233167], which performs sequence alignment at regions flanking SVs while considering large deletion and insertion blocks. Conventional global and local sequence alignment algorithms cannot handle the alignment of sequences to breakpoints.



**Figure 1. 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.

**[XXX - expand with results from 1000G]**

*Tools for complex event identification and assembly.*Although detection of deletions is becoming routine, identification and characterization of insertions, inversions, tandem duplications and translocations, as well as complex events involving various combinations of these events, is an unsolved problem. It is recognized now that these complex events are more common than previously believed (CITE). We have made progress on analyzing such events using both WGS and WES data. We developed and published PEMer [19236709] for identifying complex rearrangements from WGS datasets.

In one of the first studies of its kind, we comprehensively characterized complex SVs from a large cohort of TGCA WGS datasets [23410887] and validated them *in silico* using assembly-based methods such as TIGRA-SV. TIGRA employs a de Bruijn graph–based approach with several innovative extensions: 1) using multiple k-mers to increase the chance of assembling low-coverage alleles, 2) recording alternative paths in the contig graphs to represent the different alleles, 3) using reads (instead of k-mers) to resolve repeats in the graph, and 4) taking various measures to enhance the representation of the alternative structures. These contigs, typically several hundred bp in length, are then realigned to the human reference for determining exact breakpoint locations. Among our findings, we observed [23410887] that complex events have higher allele frequencies than simple events and perhaps happened earlier in the evolution of the disease genome than other events.

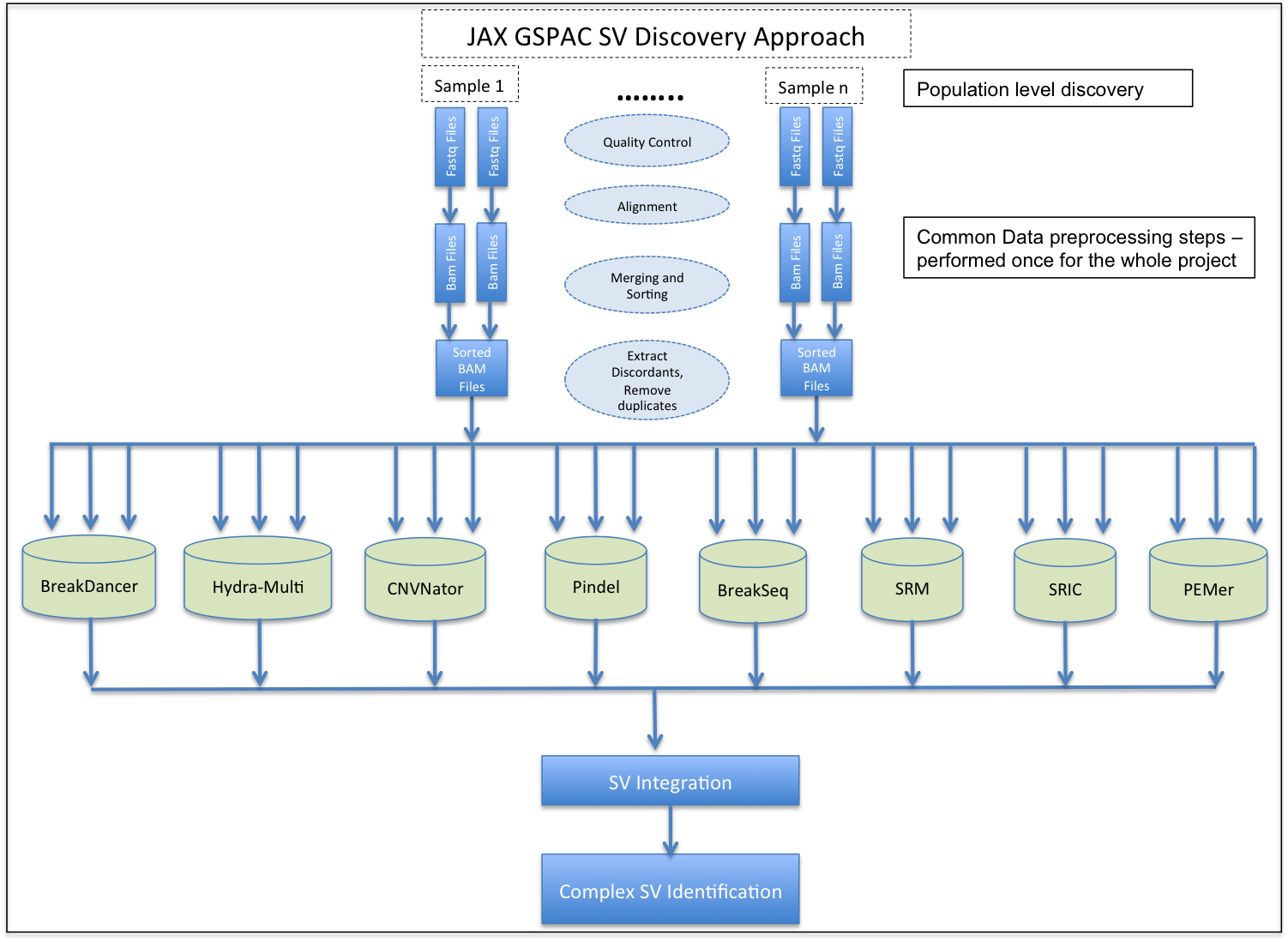
*Extensive complexity at structural variation breakpoints.* As part of the 1000GP SV analysis team, we assessed the complexity of the 29,954 deletions where the breakpoints had been sequenced and assembled. Consistent with the clustering analysis and the observed repeated rearrangement of duplication sites, 7.1% (1822) of these deletions intersected another deletion with different breakpoints. Moreover, a larger fraction (16%) of assembled deletion sites showed the presence of additional inserted sequence at deletion breakpoints. We grouped 1,651 deletions with at least 10 bp of additional DNA sequence between the original SV site boundaries into four broad classes to classify them in terms of variant complexity (Figure 2a). The most common class (N=501, 30%), termed *Ins with Dup and Del*, comprised deletions exhibiting a recognizable duplicated sequence interval within the respective inserted sequence. Notably, in many cases (N=191, 12%), the inserted sequences comprised two or more duplications of sequences outside the deletion boundaries (a class denoted Ins with *MultiDup and Del*). Furthermore, a class of complex sites that we term *MultiDELs* (N=370, 22%) consisted of two or more adjacent deletions separated by an inverted or non-inverted “spacer” (or multiple spacers) of up to ~204 bp in length. However, we note that not all sites clearly fit into the classes depicted in Figure 2a, with 214 sites forming distinct patterns corresponding to multiple classes or exhibiting increased breakpoint complexity. Our study within the 1000GP sample cohort also showed that an appreciable fraction (80%) of inversions are complex (Figure 2b), likely involving DNA replication errors [19597530, 19180184].



**Figure 2. 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 show much higher levels of complexity than expected. (taken from Phase3 paper – CITE)

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. A misidentified SV (i.e., inverted duplications classified as inversions) could completely alter the biological interpretation at that breakpoint locus. With deep coverage datasets, we now have the ability to observe these variation events at nucleotide resolution, thereby surmounting the risks associated with typical SV analysis.

***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* (integrated Analysis of Structural Variants) will deliver several features, including 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), local assembly and longer read tools? for *“in silico”* validation of the SV event.



*Development of a novel integrated and scalable methodology for discovering structural variations from thousands of genomes.*During phase 3 of the 1000 Genomes Structural Variation project, we used an ensemble of 9 different algorithms for SV discovery. The individual call-sets from each of the different algorithms were merged and unified into a single release through a detailed procedure that involved re-genotyping SV genomic loci using [[MG-comment]],,GenomeStrip with a special emphasis on genotype concordance for overlapping sites. Using an ensemble of methods allows us to avoid bias that might be present in a single method. This study also illustrates our prior experience in integrating multiple SV calling algorithms into a single pipeline.

*Pipeline for population-level structural variant discovery.* As shown in Figure 3 [XXX – Redo Fig to reduce white space], we will develop a pipeline for population-level SV discovery that has the following salient features:

1. MySQL database–based sample tracking of data files through the various steps of 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.
5. Validation for the merged set of SV sites using an assembly-based approach.
6. Complex SV identification using tools for assessing breakpoints at nucleotide resolution.

To call SVs across projects that are the primary data-producing constituents of the GSP, we plan to use our extensive experience in the calibration of heterogeneous datasets in three phases:

*Phase 1—Standardization:* The pipeline will incorporate a novel a machine learning–based approach to calibrate and test the various parameters of the different SV calling methods. We will initially focus on 100 “simulated truth” (ST) samples generated using WGSim (CITE), 50 deep coverage “known truth” (KT) samples from 1000GP SV Project (CITE) and 200 test cohort (TC) samples (selected from the TCGA consortium). The ST, KT and TC cohorts represent datasets which have some evidence for the associated SV. These datasets will be given different weights in the eventual determination of the parameters of the pipeline – 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 (~2K of the proposed 200K sequenced by the CCDGs). This cohort will be used to test our pipeline for efficiency and eventual scale up to a much larger cohort in the next discovery phase. Based on the data access and compute strategies utilized in the GSP, we will explore opportunities for massive parallelization where the tools already allow for the process to be parallelized. The various compute intensive steps in the discovery pipeline are 1) the alignment of raw reads to the genome, 2) clustering of aberrant reads, 3) SV validation using assembly and 4) SV integration. These would be the primary target for optimizations. When a step is not easily parallelizable we will explore ways to extend the existing methods to make them more efficient / parallelizable.

*Phase 3—Discovery:* The main goal of this aim is to perform robust and integrative analysis for SV discovery across several thousand WGS datasets (~5% or 10K of the proposed 200K individuals sequenced by the CMGs and CCDGs). **Selection of discovery cohort of individuals [XXX – Add, why 10000 (refer to power analysis in Aim 3), availability of phenotype data]**

*Data access strategies: The JAX CSVA Cloud*. The data footprint and computing requirements for SV discovery in this phase present unique challenges. Total storage of this subset of samples is expected to require ~4 PB based on TCGA WGS statistics. To deal with this, we propose to develop the JAX CSVA Cloud for large-scale computing. The JAX CSVA Cloud will unite the extensive data storage and computational tools present at JAX [[MG-comment]],and other institutes into a single resource and will be available to all the members of the JAX CSVA. Our two-stage local and cloud approach for analysis and data access is as follows:

*i) The JAX local data center.* In a traditional data center model, data is downloaded over the internet for analysis on local high-performance computing resources. JAX has extensive infrastructure for this model, including an HPC cluster at JAX-GM with 1700 cores and 1.4 PB of storage that JAX will further expand over time (see Facilities and Resources). While this is smaller than the size of resources required for overall number of samples to be analyzed, we can analyze the full set by transient download and analysis of raw data with retention of only necessary results. JAX has dedicated IT staff for transfer, storage, and system administration activities on the cluster, including expertise with permissions management and secure HTTPS data transfer using clients such as Aspera. Method development and initial analysis will be performed by consortium members primarily at the JAX local data center.

*ii) Cloud-based data access model.* After initial method development and analysis, we plan to disseminate methods to the broader research community using cloud approaches. 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 also evaluated the suitability of Amazon archival storage for genomics datasets. Similarly, the Washington University group has been developing a secure, HIPAA-compliant web-driven variant discovery and annotation platform called GenomeVIP (Genome Variant Investigation 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.

To stabilize the value of such computational methods for the broader research community, JAX is now evaluating multiple commercial genomics cloud service providers (CSPs) to work with to address scale and interface issues (see letters: Seven Bridges Genomics and IBM). As part of this, JAX is now recruiting an additional 1–2 full time employees for platform evaluation and piloting. These activities are proceeding independently of this U01 proposal. JAX’s express goal is to choose a platform that will allow scientists without dedicated HPC training to analyze their datasets through a well-supported 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.

**Add Validation using known sites and de-novo targeted assembly methods [XXX – expand]**

***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 (see Aims 2,3). Results from this aim will also help answer questions about complex SV formation and population-level associations of SVs across multiple studies, thereby adding important dimension and value to the datasets being generated by the GSP. Furthermore, by making the iASV pipeline available as a community resource, we expect this work will help to propel future studies in genome-level SV analysis. These outcomes would not be possible without the depth and range of expertise resident within the JAX CSVA and thus position the proposed Center as a valuable contributor to the overall goals of the GSP. We envision the CSVA, as a hub for understanding functional impact of genetic variation, would grow into a integral component of the GSP and similar future NIH endeavours into disease biology.

***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 have knowledge about what kind of phenotype data is being collected for these patients and what orthogonal data (genomic or proteomic) that shall also be available for the same patient cohorts. In order to overcome these challenges, we shall (A) leverage the extensive experience of the assembled team to handle complex datasets (see Prelim data section), (B) make the proposed SV discovery pipeline (iASV) robust to handle diverse and complex datasets thereby suitable equipped to handle any type of data generated by the GSP.

***Cross-program objectives:*** Our studies in this Aim would allow the JAX CSVA to make critical contributions to several cross program objectives. 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.

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

***Rationale.*** Complex SVs are frequently associated with genetic diseases and are responsible for more nucleotide variation in the human genome than single nucleotide polymorphisms. 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, we anticipate that comprehensive assessment of the functional impact of SVs will require the integration of large-scale data resources such as ENCODE, 1000 GP and GTEx. We also 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.

***Preliminary data.***

*Mutational mechanisms and patterns of selection of structural variants revealed by analysis of 1,092 genomes.*The sequence content of SVs, in particular the bases around their breakpoints, carry important information about SV origin. Using datasets from 1000GP, we have intensively studied the distinct features of SVs originating from different mechanisms {24092746,26028266}. The most notable type, non-allelic homologous recombination (NAHR), is associated with activating 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 also performed SV mechanism annotations for the 1000GP Phase 3 deletions using BreakSeq\ cite{20037582}, categorizing 29,774 deletions into NAHR, non-homologous recombination (NHR), transposable element insertions (TEI) and variable number tandem repeat (VNTR) by their creation mechanisms. Among these, NHR proved to be the most prevalent mechanism (~73% of all categorized deletions) [1000G Phase3 SV reference].

*Tools for assessing functional impact of genomic variation in genes and pseudogenes.* We have extensive experience in the functional interpretation of coding mutations. To this end, we developed Variant Annotation Tool (VAT, vat.gersteinlab.org) to annotate protein sequence changes of mutations. VAT provides transcript-specific annotations according to synonymous, missense, nonsense or splice-site-disrupting changes\cite{22743228}. We used VAT to systematically survey loss-of-function (LoF) variants in a cohort of 185 healthy people as part of the pilot phase of the 1000 GP\cite{22344438}, distinguishing deleterious LoF alleles from common LoF variants in nonessential genes. We have also performed an integrative annotation of variants from 1,092 humans within phase 1 of the 1000 GP\cite{24092746}. By using enrichment of rare nonsynonymous SNPs as an estimate of purifying selection, we showed that genes tolerant of LoF mutations are under the weakest selection, whereas cancer-causal genes are under the strongest. We also participated in phase 3 of the 1000GP in which we focused our efforts on analysis of LoF variants and functional impact of SVs. We found that a typical genome contains ~150 LoF variants. We 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.

Mammalian genomes also contain many genomic “fossils” i.e., pseudogenes, which are disabled copies of functional genes retained in the genome by gene duplication or retrotransposition events. Pseudogenes are important resources for understanding the evolutionary history of genes and genomes. [[MG-comment]]yesWe developed PseudoPipe, the first large-scale pipeline for genome-wide human pseudogene annotation\cite{16574694}, and then obtained “high-confidence” pseudogenes by combining computational predictions with extensive manual curation\cite{22951037,25157146}. Based on their sequence comparisons, we identified parent gene sequences from which the pseudogene arises \cite{22951037}. We also studied the mechanisms of pseudogene formation by relating pseudogenes to segmental duplications\cite{20615899} and retroduplication events\cite{24026178}. Through integration of functional genomics data generated by the ENCODE Project, we identified a broad spectrum of biological activity for pseudogenes and, in particular, revealed that ~15% of pseudogenes are transcribed\cite{25157146}.

*Tools for evaluating functional impact of variation in coding and non-coding (nc) RNAs and regulatory regions*. RSEQtools and IQseq are tools that build gene models and determine gene and isoform abundance from RNA-Seq data \cite{21134889, 22238592}. [[MG-comment]],yes Beyond quantification of RNA in gene regions, we have also been interested in identifying transcription in unannotated regions, and have developed tools to help quantify specific types of transcripts that require special processing, particularly pseudogenes and fusion transcripts \cite{17567993,25157146, 22951037, 20964841}. We have applied our expertise in RNA-Seq analysis to analyze and compare the transcriptomes of human, worm, and fly, using ENCODE and modENCODE datasets. We found a finding striking similarity between the processes regulating transcription in these three distantly related organisms \cite{21177976, 25164755, 22955620}.

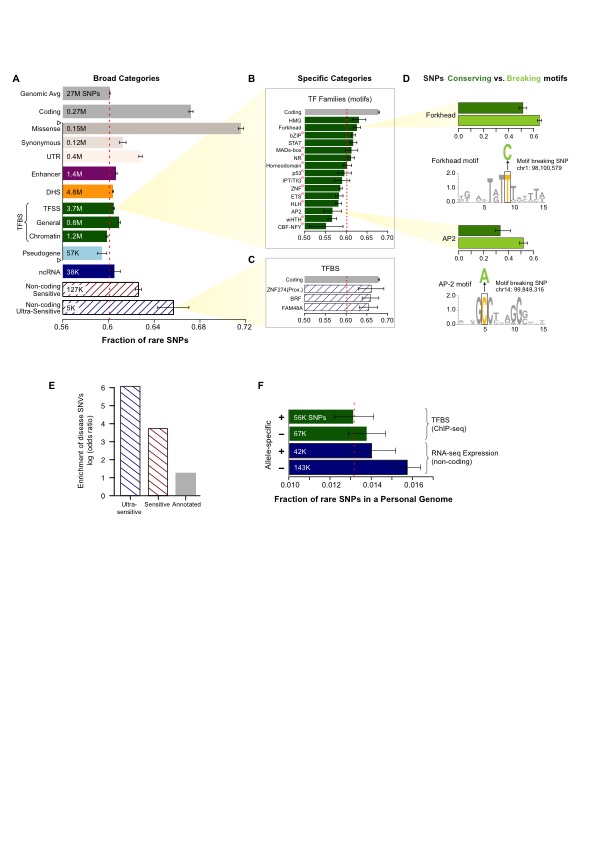
[[MG-comment]]agree We have also developed tools that specifically analyze features of ncRNAs. Our incRNA pipeline combines sequence, structural and expression features to classify newly discovered transcriptionally active regions into RNA biotypes such as miRNA, snRNA, tRNA and rRNA\cite{21177971}. 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 \cite{21596777}.

We also have extensive experience in annotation of nc regulatory regions, and in developing tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. For ChIP-Seq data we developed PeakSeq and MUSIC, which identify regions bound by transcription factors (TFs) and chemically modified histones \cite{19122651, 25292436}. PeakSeq has been widely used in consortium projects such as ENCODE \cite{19122651, ENCODE main paper}. MUSIC is a newly developed tool that uses multiscale decomposition to help identify enriched regions in cases where strict peaks are not apparent. This tool has the advantage that it robustly calls both broad and punctate peaks\cite{25292436}. We have further developed methods that use both supervised and unsupervised machine-learning techniques to identify regulatory regions (such as enhancers) and predict gene expression from ChIP-Seq data \cite{21324173, 22039215, 22955978, 25164755, 22950945}. Target Identification from Profile (TIP) can be used to predict a TF’s target genes\cite{22039215}. Furthermore, we have analyzed patterns of variation within functional nc regions, along with their coding targets\cite{21596777,22950945,22955619}. We used metrics, such as diversity and fraction of rare variants, to characterize selection pressure on various classes and subclasses of functional annotations\cite{21596777}. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region\cite{22955616}.

*Tools for network and allelic expression analyses [[MG-comment]]help annotate Func impact reveal* A powerful way to integrate diverse genomic data is through network representations. We have great experience studying regulatory networks and relating variants to networks. In particular, we have integrated multiple biological networks to investigate gene functions. We found that functionally significant and highly conserved genes tend to be more central in various networks\cite{23505346} and positioned at the top of regulatory networks \cite{22955619}. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery\cite{18077332,23505346}). Incorporating multiple network and evolutionary properties, we developed a computational method, NetSNP\cite{23505346}, to quantify the indispensability of each gene. This method shows strong potential for interpreting [[what? Phenotypic effects?]] of variants involved in Mendelian diseases and in complex disorders probed by GWAS.

We have also developed a wide range of approaches for analyzing and constructing biological networks, with a particular focus on regulatory networks. We constructed regulatory networks for data from the ENCODE and modENCODE projects, identifying functional modules and analyzing network hierarchy\cite{22955619}. To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM) \cite{25880651}. We also developed Loregic to integrate gene-expression and regulatory network data, characterize the cooperativity of regulatory factors and interrelate gate logic with other aspects of regulation, such as indirect binding *via* protein-protein interactions, feed-forward loop motifs and global regulatory hierarchy\cite{25884877}. We have also introduced several software tools for network analysis, including Topnet, tYNA and PubNet\cite{14724320, 17021160,16168087}.

Finally, we also developed a tool, AlleleSeq\cite{21811232}, 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 (SNPs, indels, and SVs).



*FunSeq : Tools for integrated functional prioritization.*We cite{24092746,25273974} recently developed a prioritization pipeline called FunSeq (Fig xxx) that identifies sensitive and ultra-sensitive genomic regions (i.e., those annotations under strong selective pressure as determined using genomes from many individuals from diverse populations). Initially developed for the prioritization of somatic mutations in cancer, FunSeq links each nc single-nucleotide mutation to target genes and prioritizes such single-nucleotide variants 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. It also detects their disruptiveness in TF binding sites (both loss-of and gain-of function events). We further developed FunSeq (FunSeq2) integrating large-scale data from various resources (including ENCODE and 1000GP) with cancer genomics data and it scores somatic recurrent mutations higher than those that are non-recurrent.[[MG-comment]]ok Using FunSeq2, we identified ~100 nc candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples\cite{24092746}. Drawing on this experience, we are co-leading the International Cancer Genome Consortium’s (ICGC) analysis of mutations in regulatory regions group.

***Research plan.***We will extend the current FunSeq prototype from its focus on somatic single-nucleotide variants 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 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 (F svtype, 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\cite{24092746}.

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.

In order to indicate the significance level of an impact score (ISorig), we will randomly shuffle the location of SVs across the genome 1000 times. We will calculate the impact score for the shuffled SV in each iteration. Finally, for the given SV, we will assign a normalized SV impact score (ISnorm) by calculating the Z-score for the original impact score compared to impact scores derived from the 1000 random shuffling of the SVs in the genome.

,

,

*Evaluating effect of structural variants on protein-coding genes*. SVs in coding regions either totally engulf an exon or gene or intersect with functional regions. The latter variants, the majority of which are[[MG-comment]] SVs or indels, usually disrupt the frame and as such result in loss of function. LoF mutations can cause potential non-sense-mediated decay and loss of important protein domains, and can alter post-translational modification sites and conserved sequences. Another concern, as shown in\cite{22344438}, is that LoF variants are prone to calling artifacts. Hence, we will focus on characterizing functionally annotated LOF variants.

Here we will further develop[[MG-comment]],vat into new tool ALoFT to substantially expand the analysis of each LoF variant with mismapping, functional, evolutionary and network features. We will quantify the confidence of LoFs using features such as whether they are in highly duplicated regions, the number of paralogs and pseudogenes, and whether they appear in the ancestral state. For functional features, we will incorporate protein structures and gene expression levels in different tissues. 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. Currently, most methods provide a dichotomous classification, i.e., benign versus disease. 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. We will investigate various machine-learning methods, and evaluate multiple independent datasets, such as mutations discovered in the CMG. This method will be the first to be developed for direct quantification of the consequences of loss-of-function mutations at the variant level.

The SVIM platform will also incorporate pseudogene analysis. Most of the time we will not expect strong functional impact for variants in pseudogenes, however pseudogenes can be source of mismapping and will be used to eliminate this confounding factor. Homologous regions such as pseudogenes give rise to a multitude of problems in variants calling. Errors due to mismapping of short reads derived from pseudogenes to genic regions leads to false variant calls. Besides, real variant calls can be missed due to reads being mapped to pseudogenes rather than the true genes\cite{25157971}. Incorporating the pseudogene annotation will improve the accuracy on quantifying LoF effects.

*Prioritizing non-coding transcripts from structural variant data.*Unlike protein-coding genes, ncRNAs lack defined open reading frames and a triplet genetic code to give clues to the impact of genetic insertions and deletions. 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 and miRNAs from publicly available data, such as CLIP-Seq, CLASH and computational predictions (TargetScan) to create a compendium of biochemical interactions with RNA\cite{25416797, 24297251, 20371350, 23622248, 21909094}. Consistent with our analysis of TF binding sites, we will define motifs that are important for RNA-binding and will assess creation and breakage of motifs by SVs. We will extend this analysis to other key motifs within RNA, such as splicing and polyadenylation sites and chemical RNA base modifications\cite{18369186}. 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 addition, we will also penalize duplications that lead to a change in the distance between the existing binding motifs within functional noncoding regions. In the prioritization scheme, we will also penalize newly created motifs if they occur close to an existing TF motif.

SV can impact TF binding sites by completely or partially deleting the TF binding motifs (motif breakers). In addition, duplication events can introduce new motifs into promoters, enhancer and other functionally important genomic elements (motif formers). Furthermore, events such as translocations and inversions can change the spacing and orientation of binding motifs within these genomic elements. We will first update the TF binding nc elements from the original FunSeq approach. Here, we will use the better enhancer definition provided by the Epigenome Roadmap \cite{25693563,25533951,25693566}, and more recently from 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. The patterns within the signal of various epigenomic features are learned using new next-generation sequencing massively parallel assays for discovering active regulatory regions in the genome. For impactful events at TF binding sites, we will use motif breakers and formers to identify SV events that are more likely to have deleterious consequences \cite{23512712, 24092746, 21596777, 23348503, 23348506, 23530248, 23887589}.

*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. The epigenetic context will be utilized to prioritize SVs: we will build tissue-specific networks and prioritize SVs impacting ubiquitous regulatory elements – protein-coding genes, ncRNAs, and TF binding sites active in multiple tissues.

*iii) Allelic activity.* We will generate a set of “allelic elements”—genomic regions featuring allele-specific behavior—by identifying allelic variants from hundreds of individuals amassed from 1000GP.\cite{23128226} Matching ChIP-seq and RNA-seq data (gEUVADIS\cite{24037378} and ENCODE\cite{22955616}) will be mapped to personal genomes constructed for each individual using their genomic sequence variants to identify variants associated with allele-specific binding and expression events, respectively. These 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.

*[[MG-comment]]move below SVIM software engineering.* We will make great efforts to make SVIM computationally efficient and able to support the large-scale computing proposed for this aim. In particular, our implementation will allow us to modularize SVIM into building a complex data context. To build the data context, we will integrate large-scale publicly available data resources, such as SVs from 1000 Genomes project \cite{23128226}, conservation data from Bejerano *et al.* and Cooper *et al.* \cite{15131266,15965027}, functional genomics data from ENCODE \cite{22955616} and Roadmap Epigenomics Mapping Consortium \cite{20944595}. We anticipate this step will be very time-consuming, as we will process large-scale genomic data into smaller summary files (e.g. associations between distal regulatory elements and likely target genes).

***Program-oriented activities.***We plan to make SVIM an easy-to-use tool. The tool will be available to other members of the program to prioritize distinct set of SVs or even indels. SVIM will be implemented as a downloadable tool, a web server, and a cloud instance.

***Expected results.***We propose to develop a software solution, called SVIM, to estimate the impact score of the SV set produced by aim1. We will prioritize these SVs and hand off to aim3 (genotype and association) to further comprehend the population impact of these variants. We plan to make the prioritization results broadly available; therefore, SVIM will incorporate the impact score into formats currently adopted, such as standards the Variant Calling Format (VCF).

***Pitfalls and alternative approaches.[[MG-comment]],agree*** The greatest pitfall of this approach is the overwhelming amount of whole genome data generated by Genome Sequencing Centers and, consequently, the number of SV to be reported. In order to overcome this, we plan to gradually process the results dividing the sequencing data into smaller analysis tiers and by focusing the initial analysis in specifics type of SVs. SVIM will 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. We further use data flow analysis to modularize the data context building. This greatly alleviates the time complexity issue when the data context needs to be partially updated in case of new incoming data or corrupted data files.

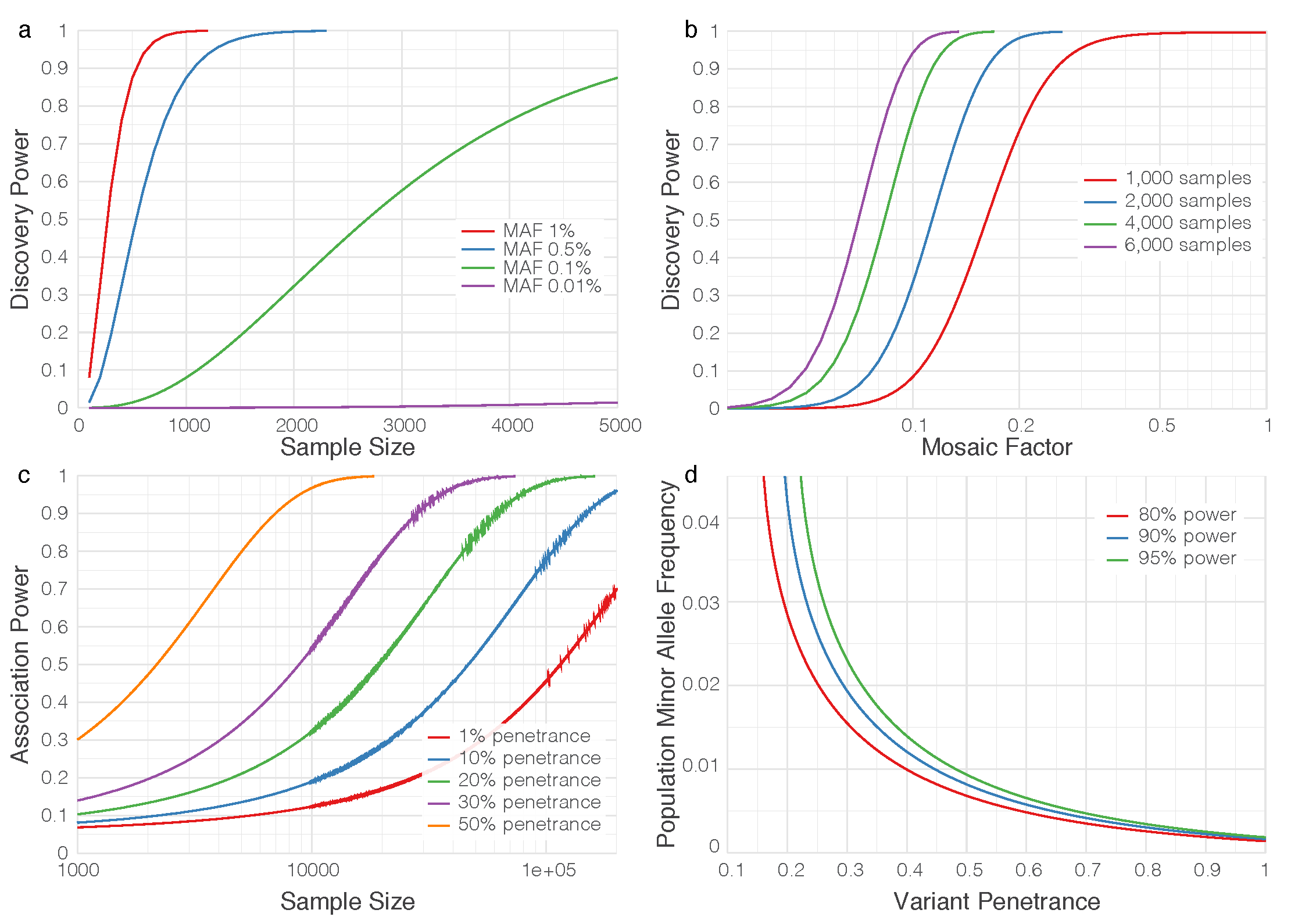
**Aim 3. Scale up to 200K samples and associating structural variants with common and rare diseases.**

***Rationale.*** We anticipate that many of the high-impact SVs we discover will be relatively rare and that no conventional tools can readily and robustly handle association analysis for SVs. We therefore envision having to develop a new association pipeline tailored to finding adequately powered SV and phenotype associations. We also anticipate that building a reference database of complex SVs in healthy individuals in Aim 1 will be essential to this undertaking.

***Preliminary Results*.**

*Power analysis for sample selection and association.* An important aspect of this study will be to select a subset of the 200K samples to be sequenced by the GSP 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 those events having the lowest population minor allele frequency (MAF) we wish to include. (Events having higher MAFs will likewise have higher discovery probabilities.) There is no general theory of discovery power covering currently used indel/SV algorithms, but we have extended an existing statistical model of coverage[1](#_ENREF_1) for simple indels to estimate the number of samples, *N*, that should be selected for full analysis. Briefly, Bernoulli probabilities for two standard discovery modes, “split reads” and “discordant read pairs”, can be derived using standard probability theory based on parameters such as read size, average insert length and variance of the library, sizes of insertion and deletion within the complex event. These parameters can be combined in various ways to model a nominal detection pipeline, e.g., “at least three observations of split or discordant reads”. Detection in each sample is binomial in the number of observations and discovery within the subset of *N* genomes is likewise binomial in the detection and MAF probabilities.

**FIGURE 3-1: Power analysis for sample selection and association.** a) Power vs number of samples, with curves for selected MAFs from 0.01% to 1%. Target events are assumed to be heterozygous and completely represented in the sample (no mosaicism). Curves are universal in the sense 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 being no mosaicism, i.e. event present in all cells; 0.5 meaning event present in half the cells in the sample, etc.), with curves for selected samples sizes from 1K to 6K. 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 equal MAF of 1% and penetrance from 1% to 50% as indicated, at the standard false-positive rate of α = 5% and 4:1 risk ratio for the Li and Leal (2008) collapsing strategy. d) curves of constant power for 10K cases/10K controls, with corresponding parameters the same as in c).



This extended model assesses *N* as a function of MAF and event size by accounting for several anticipated parameters for the data to be generated under this project, including 30X coverage per genome, average insert size of 400–600bp (20% coefficient of variation), 150bp reads, event detection based on ≥3 split reads or ≥5 discordant read pairs, and observation in at least three samples to constitute “discovery”. With these settings, the model predicts that split-reads detection will predominate for simple insertions and deletions, as well as for complex events in which one sequence segment is replaced by another. Because split-reads depend essentially only upon local alignment, power is independent of the size of events (unlike for discordant read pairs), meaning it is primarily a function of sample size and MAF. **Figure 3-1a** shows power curves for four MAFS: 1%, 0.1%, 0.05%, and 0.01%. Power at 0.1% MAF is almost 80% for 4K samples and at 0.5% MAF is ~100% for 2K samples. It drops rapidly for progressively lower MAFs, whose events are unlikely to be discovered in this study. Other types of events, like translocations, are discoverable by split-reads, suggesting their power should be comparable. Mosaicism is a potentially confounding factor in this study, for example in blood samples where an event is not uniformly present in all cells. **Figure 3-1b** estimates the effect for events at 1% MAF, showing that power is not significantly impacted until mosaicism is fairly significant. For example, it decreases by only about 10 percentage points at a mosaic factor of around 10–20% for sample sizes in the 2–4K range.

The second salient aspect of “power” is associating variants with disease phenotypes. The issues are well-known[2](#_ENREF_2), but in light of the competition between the rareness of alleles that were are seeking (down to 0.5–1% MAF) versus the very large 200K dataset, we made the following “baseline” estimates of association power. General consensus[2](#_ENREF_2) recommends “collapsing” variants for MAFs ≤1% in order to aggregate their effects for increasing power. Analysis of the widely used Li & Leal method for 10 collapsed variants (**Figure 3-1c**) shows that groupings of 1% MAF variants having high penetrance will readily associate with disease, but power will be lower as MAF and penetrance decrease. Although it is not yet known how the 200K samples will be divided over various studies, it is instructive to examine an anticipated group size of 10K cases and 10K controls (**Figure 3-1d**). Variants around 3% MAF should have ≥80% association power for penetrance of ~20% or higher while variants regardless of MAF having penetrance of <10% will likely remain ambiguous as will variants from phenotypes having substantially smaller sample allotments. Finally, in comparing discovery versus association power, it is likely we will discover more variants than what solid associations can be established for.

*Association pipeline for discovery of 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 and other relevant parameters for population/family-based association analysis. This pipeline supports popular aggregation tests, including burden tests such as the Exclusive Frequency Test (EFT)[3](#_ENREF_3), Total Frequency Test (TFT)[3](#_ENREF_3), and Cohort Allele Sum Test (CAST)[4](#_ENREF_4), and variant component tests such as the Sequence Kernel Association Test (SKAT)[5](#_ENREF_5). We have already used it to discover new associations in cohorts of thousands of samples, tailoring it to hypothesized genetic architectures of individual diseases. For example, assuming tumor suppressor genes are enriched for rare deleterious truncations, we grouped events by gene and used TFT burden testing to associate 13 genes with germline susceptibility in a cancer cohort with over 4,000 cases. In another example, it is known from common variant association analysis that some Alzheimer Disease genes carry both protective and deleterious alleles. Thus, we applied SKAT and validated the association of rare missense variants of APOE with Alzheimer’s Disease (manuscript in preparation).

**Research Plan.**

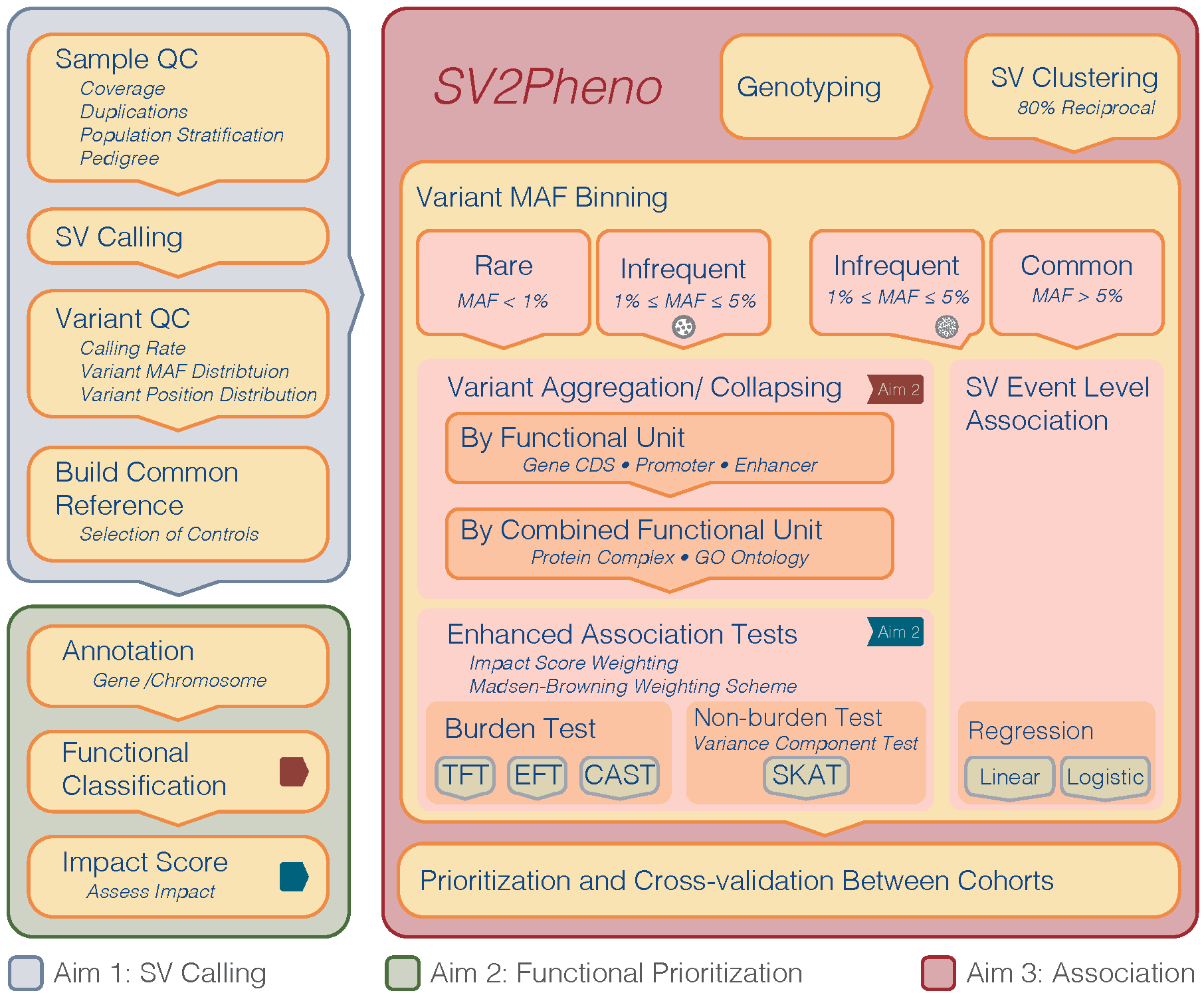
SVs are characterized by various sizes, types, penetrances and multiple alleles. We plan to genotype all SVs detected in the 3K discovery samples from Aim 1 across all 200K samples to be sequenced by the GSP over the next four years. This will allow us to obtain sufficient statistical power for genotype-phenotype associations. A critical step for inclusion of SVs in association analysis is to classify and annotate them in a meaningful fashion. By building on infrastructure and tools mentioned above, we will develop a new association analysis pipeline called “**SV2Pheno**” to infer SVs associated with phenotypes (**Figure 3-2**). Using the impact scores for each SV (Aim 2), we will further apply our strategies on cross-center and cross-platform data for integrated analysis of SNVs, indels and SVs for determining associations.

*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 for accurate determination of prevalence and allele frequencies and, importantly, will help increase the power for association analysis. This process will use **BreakSeq[6](#_ENREF_6" \o "Lam, 2010 #72)** to build a library of validated and assembled SV breakpoints and use them for genotyping individual genomes. For imprecise SVs, a combined read-pair and read-depth approach using **GenomeStrip[7](#_ENREF_7" \o "Handsaker, 2011 #73)** will be used for population-level genotyping. Conventional genotyping involves the assembly of both the reference and the alternate sequence contigs, which will be used as targets for mapping all the sequence reads present in the sample. However, given that the expected data footprint of the full set of samples is ~50PB, the traditional “*bring* *data to the computing tools”* approach will have to be upended to “*bring* *compute tools to the data”*. We anticipate that GSP, as a consortium, will develop a common data access strategy (discussed above) to enable cloud storage and computing or will make the data available through a secure ftp. In such a case, we shall build on software tools such as Sambamba (bam slicer function)[8](#_ENREF_8) that provide subsets of mappings from a large bam file. This will reduce the data footprint to a fraction of the original. As part of this process, we envision extending our TIGRA-SV assembler and Pindel split-read tools to enable them 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 more ambitious goals of this proposal:

i) Burden test power is maximized when judicious grouping results in most or all of the variants contributing to disease. However, given the various grouping options (see below), we anticipate cases in which some variants will merely be noise. *Consequently, we plan to hybridize the pipeline with more recent methods that better account for non-contributing variants*[*9*](#_ENREF_9)*.* Likewise, annotation and functional prediction can help identify irrelevant variants, which can subsequently be removed from analysis. The pipeline will receive direct feedback of this information from the ENCODE/Epigenetics Roadmap analysis.

ii) Variants are known to be associated with various diseases[10-12](#_ENREF_10), but almost certainly contribute non-uniformly; assigning appropriate weights will be necessary to wring-out maximum power. Aggregation tests can be expressed in a general way by the linear equation , where (left-to-right) is observed trait, intercept, residual, collective effect coefficient, weight of variant *i*, and tally of variant *i* (0, 1, or 2). (For example, all *wi* are set to 1 in CAST.) Here, assignment of the weights will be based on a novel combination of 3 considerations: the Madsen-Browning equation[13](#_ENREF_13) to account for allele frequency, consideration of “direction” (negative association) using e.g. aspects of the Pan-Shen approach[14](#_ENREF_14), and incorporation of our impact score (Aim 2) t o account for biological strength. In principle, such a general approach should capture signals that are too subtle for earlier-generation tests[15](#_ENREF_15).



**FIGURE 3-2: 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 under study; we will pay close attention to aspects such as population structure, sample coverage, variant call rate etc. In general, an even case-control split offers maximal power, if the total sample size is fixed. However, it is likely that the 200K sample set will furnish potentially many more controls against a given disease case set and thus some amount of power increase. In diseases where this seems feasible, we will check the available literature for any known underlying genetic commonalities and will choose extra controls in light of relevant covariates (e.g. age or smoking status).

Since we also anticipate that a high fraction of SVs will reside in nc regions, we will aggregate these 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. A criterion for what constitutes “similar” events (as it pertains to large deletions, inversions, intra-chromosomal translocations, or tandem duplications) is given by the “80% reciprocal overlap” rule,[16](#_ENREF_16) i.e., 80% of the affected base pair positions must be common to at least two samples. For large insertions and inter-chromosomal translations, we will require the breakpoints to be within 1kb of one another. We will then perform an association analysis using the collection of prototypical events and assess the significance of the associations using the impact scores generated in Aim 2.

**Level 2.** Functional Unit (Gene CDS/promoter/enhancer) level association analysis: We will perform annotation of 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). The SVs affecting a given gene will be grouped together and represented as a single, effective functional unit (**Figure 3-2**). We will then perform an association analysis using 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 perform biological annotation of the functional units in the previous step to identify any known affiliated higher-order functional 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 combined functional unit will be grouped together and represented as a single unit. We will then perform an association analysis using these units, again considering the SV impact scores generated in Aim 2. This approach has the potential to discover novel combinations of SV-containing functional units.

We will apply the tiered approach described above and association analysis approaches shown in **Figure 3-2** to analyze all genotyped samples that passed our extensive coverage and variant-calling QC from various cohorts (anticipating a number of 10–20) to identify promising candidate SVs associated with specific phenotypes.

*Integrate various types of variants (including SNVs and small simple indels) for association analysis.* The potentially most powerful analysis will come from combining information from SNVs, small simple indels and SVs for association analysis. Traditionally, weights in burden tests have been developed to account for variants with varying MAFs, but favoring variants with lower MAFs[5](#_ENREF_5),[13](#_ENREF_13" \o "Madsen, 2009 #78) such as using a parameterized beta distribution. 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 considered aggregating variants of different types. Here, we propose two methods for such integration guided by the following hypotheses: i) We hypothesize that SVs would have stronger functional impacts than missense SNVs, on average. To test this, 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. ii) We hypothesize that alterations from functional regions regardless of size contribute to phenotype. Therefore, alternatively, we plan to use SNV/indel and SV for independent burden analyses and then use standard methods based on probability theory to combine the P-values from these independent tests.

***Expected results and program-driven activities.*** This aim will culminate in the JAX CSVA **SV2Pheno** association pipeline and its associated/support tools for systematic discovering SVs associated with specific phenotype/disease. The initial version will be “packaged” (binaries, source code, makefiles, etc) and distributed for broader community use. Further, JAX CSVA will provide code support and work closely with other GSP centers to integrate SVs with SNV/indel data generated by CCDG and CMG centers. Finally, the SV reference set across ethnic groups we propose to construct will not only facilitate the discovery of significant associations by JAX CSVA, but also by the greater GSP and the broader research community.

***Pitfalls and alternative approaches.***Our expertise in power analysis (see preliminary results) renders our team well-positioned to detect SVs with MAFs between 0.5–1% using 3K 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. Specifically, there are several strategies for selecting these cases for the initial discovery: 1) from one homogenous cohort; 2) from one CCDG center across multiple cohorts; 3) from multiple cohorts generated by multiple CCDG/CMG 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 association analysis may require extensive tuning and we will use reported SVs associated with diseases[10-12](#_ENREF_10) as positive controls for adjusting impact score based weighting.

**Reference**

1 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).

2 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).

3 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).

4 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).

5 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).

6 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).

7 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).

8 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).

9 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).

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

11 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).

12 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).

13 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).

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

15 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).

16 Chen, K. et al. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome research 24, 310-317, doi:10.1101/gr.162883.113 (2014).