RESEARCH STRATEGY

SIGNIFICANCE

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

We are now compiling vital whole-genome data that will form the basis for comprehensive analyses of human genetic variation and will address current gaps in our understanding of complex diseases. In many disease contexts, known common single nucleotide variants (SNVs) account for a significant amount of phenotype variability. However, given that SVs are common, larger in size and highly structurally diverse, they are also 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. At present, most studies do not capture the spectrum of complex SVs present in genomes, and therefore this complexity is not adequately accounted for in disease association studies. Furthermore, the functional impact of SVs, especially in noncoding regions, has not been investigated systematically. Surmounting these issues will depend on novel computational methodologies for i) mining whole genome sequencing 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.

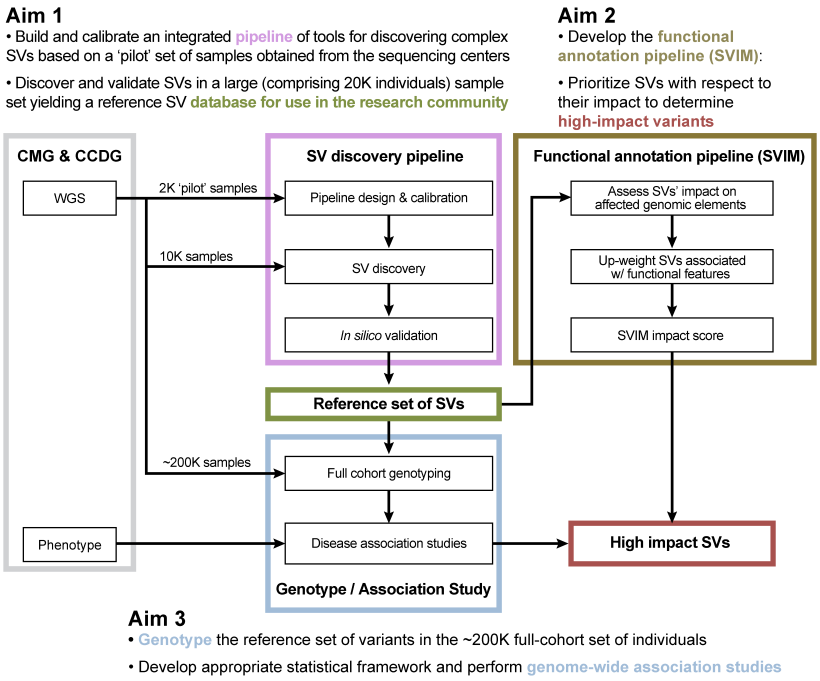

Figure 1. Overall research plan for the JAX CSVA.

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 integrate RNA-seq data and 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 the 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 ~200,000 samples of the GSP, which will provide the necessary statistical power for meaningful genotype-phenotype associations for disease-based SV association studies (Aim 3). We will be able 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 sequencing 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 the proper 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 current 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 SVs with functional data from coding and non-coding (nc) parts of the genome, especially through the integration of RNA-seq data. **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.**

RESEARCH STRATEGY

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

Rationale. SVs account for more nucleotide variation in the human genome than SNPs and therefore are likely to be associated with many genetic diseases. However, little is still known about their functional impact at a genome-wide level. SVs 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. This proposal will catalogue the largest number of SVs so far and, more importantly, integrate RNA-seq data to functionally prioritize SVs in preparation for disease association studies.

Preliminary data.

Tools for assessing functional impact of genomic variation in genes and pseudogenes. We developed Variant Annotation Tool (VAT) to annotate the impact of protein sequence mutations. VAT provides transcript-specific annotations of mutations according to synonymous, missense, nonsense or splice-site-disrupting changes[28]. We annotated variants from 1,092 humans in Phase 1 of the 1000GP[25] and observed that genes tolerant of loss-of-function (LoF) mutations are under the weakest selection and cancer-causal genes under the strongest selection. In 1000GP Phase 3, we found that a typical genome contains ~150 LoF variants and discovered significant depletion of SVs (including deletions, duplications, inversions and multiallelic CNVs) in the coding sequences, untranslated regions and introns of genes compared to a random background model, implying strong purifying selection.

Tools for evaluating functional impact of variation in non-coding (nc) RNAs and regulatory regions. We developed tools to specifically analyze ncRNAs. Our incRNA pipeline combines sequence, structural and expression features to classify newly discovered transcriptionally active regions into RNA biotypes such as miRNA, snRNA, tRNA and rRNA[29]. 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].

To better understand 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[31,32]. PeakSeq has been widely used in consortium projects such as ENCODE[31,33]. 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]. 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]. 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 37,38]. We used metrics, such as diversity and fraction of rare variants, to characterize selection pressure on various classes and subclasses of functional annotations[30]. We have also defined variants that are disruptive to a TF-binding motif in a regulatory region[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] and are positioned at the top of regulatory networks[38]. Further studies showed relationships between selection and protein network topology (e.g., quantifying selection in hubs relative to proteins on the network periphery[39,40]). Incorporating multiple network and evolutionary properties, we developed NetSNP[39] 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]. To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM[41]).

FunSeq: Tools for integrated functional prioritization. We recently developed a prioritization pipeline called FunSeq[25,43] 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 based on scaled network connectivity. FunSeq identifies deleterious variants in many nc functional elements, including TF binding sites, enhancer elements and regions of open chromatin corresponding to DNase I hypersensitive sites, and detects their disruptiveness in TF-binding sites (both LoF and gain-of-function events). We further enhanced FunSeq (FunSeq2) and identified ~100 nc candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples[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] 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.

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[25,26]. For example, non-allelic homologous recombination (NAHR), is associated with active enhancers and an open chromatin environment. Our analysis also showed that micro-insertions, flanking non-homologous breakpoints, originate from late-replicating genome loci with characteristic distances from breakpoints. We further performed SV mechanism annotations for the 1000GP Phase 3 deletions using BreakSeq[27], categorizing 29,774 deletions by their creation mechanisms. Among these, NHR proved to be the most prevalent mechanism (~73% of all categorized deletions)[17]. These results inform us on the molecular mechanisms underlying SV formation and also indicate differences in functional impacts of different SV types.

Tools for uniform processing of RNA-seq data. We have considerable expertise in analyzing RNA-Seq data, including experience in developing and setting up pipelines for the processing of RNA-seq data; specially for long RNA-seq data for ENCODE, long and short RNA-seq data for the Brainspan project as well as a custom pipeline developed for the analysis of small exRNA-seq data

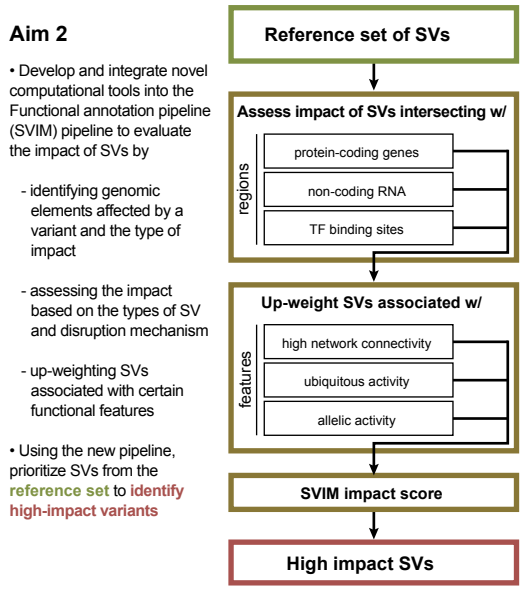

Figure 6. Overview of the functional prioritization and annotation pipeline.

for the Extracellular RNA Communication Consortium (ERCC). We have already developed an efficient in-house data processing workflow for RNA-seq data that includes data organization, format conversion, and quality assessment. RSeqTools\cite{} is a modular tool developed for the processing of RNA-seq data and generating either transcript, gene or exon level quantifications. Our lab has also developed IQSeq \cite{} which calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-seq data using a fast algorithm based on the Fisher information matrix. Another tool we developed called FusionSeq\cite{} was to detect fusion transcript in RNA-seq data, which can be important biomarker for diseases such as various types of cancer and mental diseases.

We have also developed tools specifically for linking gene expression variation to genotype, including our Allele-Seq pipeline, which quantifies allele-specific gene expression by mapping reads onto a diploid personal genome built from called genetic variants, including SNPs, short indels, and structural variants \cite{21811232}. We recently applied this pipeline on a population scale to RNA-Seq data from the 1000 Genomes Project, and used this analysis to create AlleleDB, a database of genomic regions with high allelic activity\cite{27089393}.

Research plan. To enable identification of SVs with high functional impact, we will extend FunSeq/FunSeq2 within a new pipeline called SVIM (Structural Variation IMpact)(**Figure 6)**. We will evaluate the impact score for each SV identified in Aim 1, taking into account the functional annotation of the affected genomic region and the fraction of functional elements (i.e., genes, ncRNAs and nc regulatory elements) overlapped by the SV. The impact score will also depend upon SV type (i.e., deletion, duplication, inversion or translocation).

For a given SV belonging to a particular SV type, we will 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, nc region and TF binding site) and assign relative scores to them (S_{coding}, $S_{non\text{-}coding}$, S_{TFBS}), which will vary for different SV types. Relative scores F and S will be defined for class and functional elements analogous to the FunSeq2 tool[25].

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

Identifying and predicting RNA transcripts arising from SV regions. Since structural variants often create long sequences that are not present in the reference genome, we will adapt our existing AlleleSeq pipeline to annotate the transcripts produced at SV regions. Specifically, we perform a stringent alignment of RNA-Seq reads just to an appropriately built sequence of the SV-containing regions, and then apply both our RSEQtools package \cite{21134889} and Cufflinks \cite{20436464} to generate predicted transcript models. We will then identify the most probable protein products arising from these sites and classify these as wild-type, mutant, fusion, or novel proteins to help with our downstream variant prioritization.

In addition to identifying RNA transcripts expressed from SV-containing regions from blood RNA-Seq, we will predict tissue-specific transcript models, using state of the art software for prediction of RNA splicing \cite{26496609, 25525159} and polyadenylation \cite{27095026}. We will use these transcript models as well for downstream functional analysis.

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

In addition to mutations that cause clear loss of function, we will use existing tools to prioritize SVs that we predict to cause small insertions or deletions to protein sequences. To do this, we will employ tools such as KGGSeq \cite{22241780}. [[MRS: what are the right tools to mention here]]
[[MRS: do we have anything to say about fusion transcripts, or prioritization of small deletions in proteins?]]

Prioritizing non-coding transcripts from structural variant data. To prioritize the effects of SVs in ncRNAs, we will focus on overlaps with regulatory elements and other functional regions. To perform this analysis, we will define categories of RNA regions that display human population-level conservation, and combine these features to generate RNA element scores. We will mine RNA interactions between proteins (e.g., CLIP-Seq) and miRNAs (e.g., TargetScan) to create a compendium of biochemical interactions with RNA[45-49]. We will further investigate RNA secondary structure, looking for structured regions that are highly sensitive to mutation. For these regions, we will assess deleteriousness of mutations by differences in predicted free energy or structure ensembles[50] relative to wild type. We have found annotations of all of the above types—biochemical interactions, regulatory motifs, and structured regions—that are enriched for rare variants in the human population and will use these sensitive RNA regions to score and prioritize potential deleterious SVs in ncRNA. Large SVs will ultimately be scored based on the highest scoring subregion disrupted (or created) by the SV..

Prioritizing non-coding regulatory elements from structural variant data. Unlike protein-coding genes and ncRNAs, TF binding motifs are relatively small in size. Thus, we are going to analyze duplications that occur close to these motifs and analyze where these duplications lead to the breakage of existing or creation of new motifs. In the prioritization scheme, we will also penalize changes in distance between motifs and newly created motifs if they occur close to an existing TF motif. We will first update the TF binding nc elements using better enhancer definitions provided by the Epigenome Roadmap[51-53] and ENCODE. We will further 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 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 an epigenetic context to identify tissue-specific phenotypic effects that are strongly influenced by SVs. We will prioritize SVs impacting genes, ncRNAs, and TF binding sites active in multiple tissues.
iii) Allelic activity. 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 develop a prioritization scheme for SVs overlapping these allelic elements.

[[DC]]Integration of SV annotation and RNA-seq data. The functional interpretation of SVs may partially be obtained solely by determining the genome annotations in which SVs lie. However, genome annotations alone (such as knowledge that an SV falls within a region of open chromatin) often provide limited knowledge in terms of linking SVs to phenotypic traits. To predict the phenotypic effects of individual SVs on measurable phenotypes, we will link SVs to the specific genes that they affect by performing comprehensive genome-wide searches for expression quantitative trait loci (eQTLs). A given gene may be influenced by proximal or distant eQTLs. Whereas proximal regulatory regions (such as promoters) are localized to the regions around a gene, trans-acting distant regulatory elements (such as enhancers) may be more diffuse, cover wider swaths of the genome, and act in greater multiplicity on a given gene. As such, eQTLs that result from a *specific* SNV within a distal regulatory element may be substantially weaker than those that result from an SNV within a promoter.

Furthermore, it may be expected that smaller genomic perturbations (such as SNVs) generally induce smaller effects on transcription (consistent with this hypothesis, it has previously been determined that SNVs have a reduced tendency to affect gene expressions relative to indels; \cite 24037378). Compounding these challenges associated with linking a given distal SNV to a particular gene, the stringent significance criteria required to correct for the many tests needed to identify distant trans-eQTLs render such eQTLs far more difficult to identify as a result of reduced statistical power. Together, these phenomena may make large SVs more suitable candidates in terms of identifying trans-eQTLs. Our search for SV-induced eQTLs will be accomplished by performing genome-wide searches for patterns in which the presence or absence of the SVs (identified in Aim 1) strongly correlate with the expression levels (as measured by mRNA abundance using RNA-seq data) of a battery of genes throughout the genome. Of particular interest will be those genes previously implicated in disease-associated pathways and network modules. [[FN: To calculate the eQTL, we need to genotype the SV in 10k-100k individuals with RNA-seq – which is produced only in aim 3. Should we consider moving the genotype up (from aim3 to aim2) or moving this session down (from aim2 to aim3)?]]

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

Pitfalls and alternative approaches. We anticipate that the greatest pitfalls are (i) possibly an 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 into specifics type of SVs. SVIM will also be based on the data context to efficiently prioritize variants from some WES datasets, but optimally 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 the 1000 GP[54], conservation data from Bejerano *et al.*[55] and Cooper *et al.*[56], functional genomics data from ENCODE[33] and Roadmap Epigenomics Mapping Consortium[57].

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 ensures 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 SVs 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 studies 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[80] to assess the completeness of SV census across populations and disease types. We will 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 will choose appropriate controls for our SV discovery analysis based on population structure and other confounding factors. These sets of controls will be shared across the GSP and will 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. *Curr Protoc Bioinformatics* **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 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).

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 Ouyang, Z., Snyder, M. P. & Chang, H. Y. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome research* **23**, 377-387, doi:10.1101/gr.138545.112 (2013).

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

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

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

54 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).

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

56 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).

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

58 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).

59 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).

60 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).

61 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).

62 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).

63 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).

64 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).
65 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).
66 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).
67 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).
68 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).
69 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).
70 Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223, doi:10.1038/nature12141 (2013).
71 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).
72 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).
73 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).
74 Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449, doi:10.1126/science.1138659 (2007).
75 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).
76 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).
77 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).
78 Pan, W. & Shen, X. Adaptive tests for association analysis of rare variants. *Genetic epidemiology* **35**, 381-388, doi:10.1002/gepi.20586 (2011).
79 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).
80 Schnabel, Z. E. The Estimation of Total Fish Population of a Lake. *American Mathematical Monthly* **45**, 348-352 (1938).