Specific Aims:

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

**C.2. FUNCTIONAL PRIORITIZATION**

**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].

**C.2.2.2 Functional Impact in Genes and Pseudogenes**

We have extensive experience in functional interpretation of coding mutations. To this end, we develop Variant Annotation Tool (VAT, vat.gersteinlab.org) to annotate protein sequence changes of mutations. VAT provides transcript-specific annotations and annotates mutations as synonymous, missense, nonsense or splice-site disrupting changes\cite{22743228}. We have 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 Genomes Project\cite{22344438}, distinguishing deleterious LoF alleles from common LoF variants in nonessential genes. We have done an integrative annotation of variants from 1092 humans from the 1000 Genomes Project Phase 1 study\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 have also participated in the 1000 Genomes Project Phase 3 studies on LoF variants and functional impact of SVs and found that a typical genome contains ~150 LoF variants. Furthermore, we discovered a significant depletion of SVs (including deletions, duplications, inversions and multiallelic copy number variants) in CDS, UTRs and introns of genes, compared to a random background model, which implies strong purifying selection.

We developed PseudoPipe, the first large-scale pipeline for genome wide human pseudogene annotation\cite{16574694}, and then obtained the “high confidence” pseudogenes by combining computational predictions with extensive manual curation\cite{22951037,25157146}. We identified parent gene sequence from which the pseudogene arises based on their sequence comparisons\cite{22951037}. We have 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 ~15% of pseudogenes are transcribed\cite{25157146}.

**C.2.2.3 Functional Impact in non-coding RNAs**

We have also developed RSEQtools and IQseq, tools that build gene models and determine gene- and isoform-level RNA-Seq quantifications \cite{21134889, 22238592}. Beyond quantification of RNA in gene regions, we have also been interested in identifying transcription in unannotated regions, and have developed specific 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 distant organisms \cite{21177976, 25164755, 22955620}. 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}.

**C.2.2.4 Functional Impact in non-coding regulatory regions**

We have extensive experience performing annotation of non-coding regulatory regions, with expertise in developing tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. For ChIP-Seq, we have developed two tools - PeakSeq and MUSIC - that identify regions bound by transcription factors 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 to use ChIP-Seq signals to identify regulatory regions such as enhancers and to predict gene expression, using both supervised and unsupervised machine learning techniques \cite{21324173, 22039215, 22955978, 25164755, 22950945}. We developed method called Target Identification from Profile (TIP) to predict a TF’s target genes\cite{22039215}. Furthermore, we have analyzed the patterns of variation within functional noncoding 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}.

**C.2.2.5 Preliminary results related to networks and allelic expression**

A powerful way to integrate diverse genomic data is through networks representations. We have great experience studying regulatory network and relating variants to networks. In particular, we have integrated multiple biological networks to investigated gene functions. We found that functionally significant and highly conserved genes tend to be more central in various networks\cite{23505346} and positioned on the top level 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 have developed a computational method - NetSNP\cite{23505346} to quantify the indispensability of each gene. This method shows strong potential for interpretation of variants involved in Mendelian diseases and in complex disorders probed by genome-wide association studies.

We have also developed a wide range of analyses on 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) to infer the hierarchy of a directed network\cite{25880651}. We also developed Loregic to integrating gene expression and regulatory network data and 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}.

We have also developed a tool, AlleleSeq\cite{21811232}, for the detection of candidate variants associated with allele specific binding (ASB) and allele specific expression (ASE) based on the construction of a personal diploid genome sequence (and corresponding personalized gene annotation) using genomic sequence variants (SNPs, indels, and SVs).

**C.2.2.6 Preliminary results related providing and integrated functional prioritization score**

In recent studies\cite{24092746,25273974}, we have integrated and extended methods to develop a prioritization pipeline called FunSeq (Fig xxx).This pipeline was mostly focused on prioritization of somatic mutations in cancer. It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each non-coding single nucleotide mutation to target genes, and prioritizes such single nucleotide variants based on scaled network connectivity. It identifies deleterious variants in many non-coding 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 The 1000 Genomes Project) with cancer genomics data and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq2, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples\cite{24092746}. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

**C.2.3 Research plan**

We will extend the current FunSeq prototype from its focus on somatic single nucleotide variants to allow 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 in to account the functional annotation of the affected genomic region, as well as the fraction of functional element (i.e. genes, noncoding RNAs and non-coding regulatory elements) overlapped by the SV. Furthermore, the impact score will be also dependent upon the 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 TFBS) 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.

,

,

**C.2.3.1** **Identifying the effect of SVs on protein coding genes**

SVs in coding regions either totally engulf an exon or gene, or intersect with functional regions. The later variants usually disrupt the frame and are termed LOF (loss-of-function) variants. In fact, the majority of LOF variants are SVs or indels. LOF mutations can cause potential non-sense-mediated decay, loss of important protein domains, altered post-translational modification sites and conserved sequences. Another concern about LoF variants is potential calling errors. As shown in\cite{22344438}, LoF variants are prone to calling artifacts. Hence, we will focus on characterizing functionally annotated LOF variants.

Here we will further develop 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 consisting of 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 with multiple independent datasets, such as mutations discovered in the CMG (Center for Mendelian Genomics). This method will be the first method developed to directly quantify consequences of loss-of-function mutation at the variant level.

In our analysis we will also incorporate pseudogenes. 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.

**2.3.2. Consistently prioritizing non-coding transcripts from SV data**

Unlike protein-coding genes, noncoding RNAs 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. In order to perform this analysis 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 transcription factor 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.

**2.3.3. Consistently prioritizing non-coding regulatory elements from SV 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 bindings sites (TFBS) 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 non-coding 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 learnt 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}.

**2.3.4. 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.

(1) 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.

(2) 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.

(3) 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 The 1000 Genomes Project.\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.

**2.3.5. SVIM software engineering**

In order to support the large-scale compute we will make great efforts to make SVIM computationally efficient. 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).

**2.4. 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.

**2.5. 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).

**2.6. Pitfalls and alternative approaches**

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 whole genome sequencing dataset. 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.