Specific Aims:

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 a SV can affect a genomic elements (e.g. partial overlap v engulf) and will integrate conservation information, existing genomic annotations, 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 strong allelic activity -- i.e. demonstrated functional sensitivity to variants.

A. Significance
B. Innovation
[[MG: we need to 3 lines of innovation here]]

C. Aims
C.1 SV FINDING [4pg, Charles, Ankit]
[[...]]

C.2. FUNCTIONAL PRIORITIZATION [4pg]
We will develope innovative tools to analyze their functional impacts of complex SVs[[MG: use SVs instead structural variants]]. These complex events are disproportionately in the noncoding part of the genome and we anticipate that most of functional impacts would involve integration with large-scale data resources such as ENCODE and GTEx.

C.2.1. Preliminary Results
C.2.1.1 Preliminary Results related to Mechanism Classification (NAHR v NHR)
The Gerstein lab has intensively studied the distinct features of SVs originated from different mechanisms. This indicates unique forming processes and potentially divergent functional impacts[24092746][26028266]. The most notable type, NAHR, is associated with activating enhancers and open chromatin environment. Our analysis also showed that micro-insertions flanking NH type breakpoints are templated from late replicating genome sited with characteristic distances from breakpoints. These results not only shed light on SV forming processes but also indicate differences in functional impacts of different SVs types. The Gerstein lab has also performed SV mechanism annotations for the 1000 Genomes Phase 3 deletions using BreakSeq [20037582]. We categorized 29,774 deletions into NAHR, NHR, TEI and VNTR by their origination mechanisms. Among these, NHR is the most prevalent mechanism (~73% of all categorized deletions) [1000G Phase3 SV].?

C.2.1.2 Preliminary Results related to Functional Impact in Coding Regions
[[MG: we need to update to incl. p1 & p3 + other papers]]
We have extensive experience in 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 and annotates mutations as synonymous, missense, nonsense or splice-site disrupting changes \cite{22743228}. We have used VAT to systematically survey LoF variants in a cohort of 180 healthy people as part of the Pilot Phase of the 1000 Genomes project \cite{22344438}, distinguishing LoF-containing recessive genes from benign LOF-containing genes. [[MG: need to move this elsewhere]] In this grant, we will substantially expand this analysis by developing methods that will (1) provide variant-

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: classification of the variants

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: and complex indels

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Font color: Custom Color(RGB(34,34,34)), Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,,

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: 2.2. Analyzing the functional impact [2.75pg] ... [1]

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved down [1]: 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}. Incorporating multiple network and evolutionary properties, we have developed a computational me ... [2]

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ... [3]

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,updatephase1,,,,

specific functional impact scores, and (2) distinguish between recessive, dominant and benign variations. 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.

*** prelim res for pseu

To identify pseudogenes in the human genome, we developed PseudoPipe, the first large-scale pipeline for genome wide human pseudogene annotation\cite{16574694}. We also obtained the "high confidence" pseudogenes by combining computational predictions with extensive manual curation\cite{22951037,25157146}, and identified parent gene sequence from which the pseudogene arises based on their sequence comparisons\cite{22951037}.

*** prelim res rel to ncRNAs
[MG: need to demark DART & update!]

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, a problem which we have developed a series of tools to address this, e.g. Database of of Annotated Regions and Tools (DART) \cite{17567993}. We have also developed specific tools to help quantify specific types of transcripts that require special processing, particularly pseudogenes and fusion transcripts \cite{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}.

*** prelim res rel to TFBS

We have extensive experience performing non-coding genome annotation, with expertise 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}.

** Prelim results rel. to networks

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved (insertion) [2]

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved (insertion) [3]

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved (insertion) [4]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: and RNA-Seq

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: . ... [4]

A powerful way to integrate diverse genomic data is through representations as networks. We have great expr. 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}. 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 its strong potential for interpretation of variants involved in Mendelian diseases and in complex disorders probed by genome-wide association studies.

We have 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}. We have also introduced several software tools for network analysis, including Topnet, tYNA and PubNet \cite{14724320, 17021160, 16168087}.

[MG: need to integrate this:] 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}).

[MG: add loregic,Hines]

** prelim res. related to cons. in noncoding regions

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets\cite{21596777,22950945,22955619}. We used metrics, such as diversity and fraction of rare variants, to characterize selection 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}.

*** Prelim results related to element wise burden analysis

We also have extensive experience in methods to identify significant coding and noncoding mutation hot-spots in cancer samples. The sharp decrease in sequencing costs enabled studies of large cancer cohorts and, consequently, the description of an overwhelming amount of variants in these genomes. Therefore, we developed an integrative framework for Large-scale Analysis of Recurrent Variants in noncoding Annotations named LARVA to test the somatic mutation burden on the noncoding annotated regions. For the coding regions analysis, we also developed methods to comprehensive analyse mutational using standardized sequence-based inputs along with multiple types of clinical data to establish correlations among mutation sites, affected genes and pathways, and to ultimately separate the commonly abundant passenger mutations from the truly significant events. This pipeline named Mutational Significance in Cancer (MuSiC)[22759861]. MuSiC was applied to a vast number of cancer samples including 316 ovarian cancer samples from the TCGA ovarian cancer project.

*** prelim res. related to funseq
[MG: need to integrate heading & remove redund in blocw]

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved (insertion) [5]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,,ncrna

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved (insertion) [1]

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved up [3]: 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, a problem which we have developed a series of tools to address this, e.g. Database of of Annotated Regions and Tools (DART) \cite{17567993

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved up [4]: We have also dev ... [5]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,merge w above

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved up [5]: A powerful way to ... [6]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: }.,,,,cut upd,,,

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted:

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: }.,,,,

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: Further studies showed ... [7]

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved (insertion) [6]

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Indent: First line: 0.5"

In recent studies\cite{24092746,25273974}, we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (Fig xxx). 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 mutation to target genes, and prioritizes such 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). 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 FunSeq, 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 2.2. Research plan

We will extend the current FunSeq prototype from its focus on somatic variants to allow the identification of key structural variations associated with high functional impact (Fig 3). Our new pipeline is called SVIM (Structural Variation IMpact). It will have several features tailoring it to structural variation impact analysis, including:

(1) Identifying the effect of SVs on protein coding genes;
(2) Identifying impact of SVs in the non-coding genome [ncRNA];
(3) Identifying impact of SVs in the non-coding genome [TFBS];
(4) Upweighting variants related to to allelic activity and network connectivity.

*** Identifying the effect of SVs on protein coding genes

[IMG: need to expand logic!] SVs in coding regions either totally effect an exon or gene or intersect. The later variants usually disrupt the frame and are LOF variants. In fact, the majority of LOF variants are SVs or indels [IMG: more!]. Hence, here we we will focus on characterizing SV LOF variants.

We are going to further develop our prioritization in relation to loss-of-function (LOF) mutations. LOF mutations can cause potential non-sense-mediated decay, loss-of important protein domains, post-translational modification sites and conserved sequences. Another concern about LoF variants are potential calling errors. As shown in \cite{22344438}, LoF variants are prone to calling artifacts. To quality filter and functionally annotate LoFs, we will develop ALoFT to annotate 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 they appear in, 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. 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.

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,,

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: We have also applied our method to investigate non-coding mutation patterns in subtypes of cancer \cite{submitted}.

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Left

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,cons n recurr --- ultra ,,,,burden analysis

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved up [6]:

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: 3

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: and MuSiC

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: their

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: Relating this

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: 2.3.1 !,,,,,prioritizing

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Font:Not Bold, Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: bar,,,,from structural variation data

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,,SV in coding usually log.....Weill ... [8]

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

[[MG: add in logic & perhaps put in prelim res.]] One issue with coding regions is the issue of mis-mapping due to pgenes...

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. On the other hand, real variant calls can be missed due to reads being mapped to pseudogenes rather than the true genes\cite{25157971}.

[[MG: add in logic & perhaps put in prelim res.]]We will take into accoutn pgenes

2.3.2. Consistently prioritizing non-coding transcripts from structural variation data

[[MG: add in logic]] For ncRNAs, there's no triplet and read deablement due to indels, so ... look for SV hit conserved bit also whole v part

We will searching for RNA regulatory regions that are sensitive to mutation in the human population. Specifically, we will mine RNA interactions with proteins/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}. In addition, we will incorporate aspects of RNA secondary structure, and also consider key motifs within RNAs, such as splicing and polyadenylation sites. We have found annotations of all of the above types that are enriched for rare variants in the human population and will use these sensitive RNA regions in an analogous way to the FunSeq2 tool to score and prioritize potential deleterious variants in noncoding RNA.

[[MG: expand]]

2.3.3. Consistently prioritizing non-coding regulatory elements from structural variation data

[[MG: add in logic]] Unlike protein coding genes & ncrnas, TF binding sites are small .... thus here we mostly analysis about dup nearby or far ... new site and break

Structural variations 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. In particular, 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. For impactful events at TF binding sites, we will use motif breakers and formers to identify structural variation

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,,caHomologous

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved up [2]: To identify pseudogenes in the human genome, we developed PseudoPipe, the first large-scale pipeline for genome wide human pseudogene annotation\cite{16574694}. We also obtained the "high confidence" pseudogenes by combining computational predictions with extensive manual curation\cite{22951037,25157146}, and identified parent gene sequence from which the pseudogene arises based ... [9]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,prelim,,,

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: }.,,,we will

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: L use

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted ... [10]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,,nc RNA

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: . Look

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: cons

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: We will prioritize genetic ... [11]

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: [[MRS2MG (31 July) cut ... [12]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: In order to define structu ... [13]

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: . ,,,,unlike ncrna... tf

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ...

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,deletename,,,

events that are more likely to have deleterious consequences \cite{23512712,24092746,21596777,23348503,23348506,23530248,23887589}. In a way that is consistent with our means of searching for motif-breaking variants in TF binding sites, we will identify motif-breakers in specific RNA binding motifs. Studies of RNA processing and function have identified key motifs associated with events ranging from RNA splicing to chemical RNA base modifications\cite{18369186}. For miRNA/protein bindings sites, we will likewise use the specific binding sites of the microRNAs and whether the respective structural variations move closer to or further from the canonical pattern.

2.3.4. Further Variant prioritization based on networks, tissue specificity & allelic activity

[JMG: add in logic?] We will upweight variants in the hubs of networks ...

We will evaluate the impact of SVs in different epigenetic context, as this could help identify tissue-specific phenotypic effects that are strongly influenced by SVs.This epigenetic context will be further utilized to prioritize SVs. In particular, we will build a tissue-specific networks (based on proteins that are expressed in the strongly influenced tissue), as well as a tissue-specific gene regulatory network. [JMG: add in logic?] We will prioritize ***ubiquitous*** active elements: genes, ncRNA & ...

Allele-specific structural variants potentially provide a most direct readout of the functional impact of SVs. We will prioritize SVs by quantifying their influence within 'allelic elements', or allelic regions in the genome.

We derive allelic elements by first identifying allelic variants from hundreds of individuals. These individuals will be amassed from The 1000 Genomes Project \cite{23128226}. We will match them with their corresponding RNA-Seq and ChIP-seq experiments from multiple disparate studies, such as gEUVADIS\cite{24037378} and ENCODE\cite{22955616}.

[JMG-to-TG: need to say more, map to pers genome ...&c?]

Subsequently, allelic variants (rare and common) identified across hundreds of genomes can be aggregated into ‘allelic genomic elements’. Genomic elements like exonic regions, TF binding sites and non-coding RNA showing ‘allelic activity’ will be considered as ‘allelic genomic elements’. We will prioritize ‘allelic elements’ encompassed by structural variations. More specifically, Each element will be assigned an ‘allelicity’ score based on not only its enrichment of allelic variants within the element (in comparison to accessible variants within the elements and having sufficient coverage to make an allelic activity call), but also across the number of individuals having allelic variants in a consistent allelic direction.

2.3.5. We will modularize SVIM to handle updates to a complex data context & simultaneously carry out efficient production runs

[JMG: add in logic?] As we anticipate large scale compute, we will make great efforts to make SVIM computationally efficient. In particular, our implementation will allow us to modularize SVIM into two components: (#1) building a complex data context and (#2) an efficient and high-throughput production run. To build the data context (#1), we will integrate large-scale publicly available data resources, such as structural variations 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

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved (insertion) [7]

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,move up,,,,,

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,net n prior ubiq protein-protein interaction network

MG-Sun-night-edit 8/3/2015 12:16 AM
Moved up [7]: 2.3.4.

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: Variant prioritization based on allelic activity

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,tg

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ,,,,as

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Font:Not Bold, Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: ...

MG-Sun-night-edit 8/3/2015 12:16 AM
Formatted: Font:Not Bold, Highlight

MG-Sun-night-edit 8/3/2015 12:16 AM
Deleted: effic ... [14]

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). The production run (#2) will prioritize variants from WGS based on the data context. The variant prioritization step needs to be quite efficient, so we can tackle >1000 genomes in fairly short time. 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 plan to make SVIM an easy-to-use tool. It will be implemented as a downloadable tool, a web server, and a cloud instance.

C.3 Scaling up to 200K & Doing ASSOCIATION analysis **[2.5pg, Li]**

[[...]]