* Significance

** Much recent progress in annotating the non-coding genome, making it ripe for variant annotation

Annotating non-coding regions is essential for investigating genome evolution\cite{16987880}, for understanding important biological functions (including gene regulation and RNA processing)\cite{19148191}, and for elucidating how SNPs and structural variations may influence disease\cite{15549674}. Many projects related to annotating the noncoding genome have recently come to completion. The Encyclopedia of DNA Elements (ENCODE) Project recently provided a comprehensive catalogue covering much of the entire human genome\cite{22955616}. In addition, the model organism ENCODE (modENCODE) Project presents an extensive genomic annotation of drosophila\cite{21177974} and C. elegans\cite{21177976} and a way to relate this to human. Furthermore, large-scale mRNA and miRNA sequencing have been applied to elucidate the functional landscape of regulatory variations in the human genome\cite{24037378,20220756,20220758,24092820}. Similar efforts have been directed toward annotating human epigenomic data to investigate underlying disease mechanisms\cite{23482391}. Moreover, the important role of regulatory variants in various diseases have generated a great deal of interest in identifying and annotating the expression of Quantitative Loci linked to specific genes\cite{18597885,20369019}.

** Non-coding variants, most of which are regulatory, are significant to the study of diseases but less well studied than coding variants

Numerous studies have been conducted on the mutations to coding portions of the genome. However, comparatively less effort has been invested in the investigation of disease-related disruptions to noncoding portions of the genome. Nevertheless, a few \cite{23348506, 23348503} initial studies indicate that variants in non-coding regions of genome significantly influence the associated phenotype\cite{17185560} and are often implicated in various diseases\cite{23138309,16728641}. Much of the non-coding variation is contributed by regulatory variants, where cis- and trans-acting variation in the human genome can modulate gene expression\cite{19636342} and this gene expression variation has been implicated in cancer and other diseases\cite{23374354,23348506,23348503,7663520,19165925,18971308}. Specific examples are expression quantitative trait loci (eQTLs) and variants associated with allele-specific behavior. It has been shown that a significant fraction (26%-35%) of inter-individual differences in transcription-factor (TF) binding regions coincides with genetic variation loci and that about 5% of transcripts levels are associated with inherited variant states\cite{20299548}. Genotype-transcript associations have been reported at large for multiple types of inherited variants\cite{21479260,20220756,20220758,21862627,1728997}, however experimental evidence of inherited variants, allele-specific effect on enhancer/promoter activities and transcriptional influence (short and long range) are lacking.

* Approach

** Much Previous Experience with SV calling which is immediately applicable

We have been members of the 1000 Genomes Structural Variation group (1000SV) since its inception. In the context of this group we have developed numerous algorithms for SV detection and characterization. These can be easily applied in the Mendelian context. Moreover, there is much convincing evidence that SV while fewer in the genome tend to be higher impact variants. For instance, studies identified many rare and de-novo structural variants have been found to be associated with many diseases like diabetes\cite{25424174} and neurodegenerative diseases like autism, tourette syndrome, bipolar disorder, and schizophrenia\cite{20531469,22169095,24098143,21658582,20368508,21358714}

CNVnator. // We have developed a tool, CNVnator for CNV discovery and genotyping \cite{21324876}. CNVnator utilizes mean-shift approach along with GC correction and novel multiple bandwidth partitioning to identify wide ranges of CNV events. We have calibrated CNVnator by leveraging extensive validation exercise performed by the 1000 Genome Project. Furthermore, CNVnator can detect CNVs and provide genotype information on a population level. In addition, CNVnator can also accurately detect atypical CNVs including de novo and multi-allelic events.

PEMer // In addition to copy number variants, we also developed algorithms for identification of complex rearrangements. On this end, we developed PEMer\cite{19236709} , which identifies clusters of discordantly mapped reads and compares with a simulated background model for scoring to identify complex rearrangement breakpoints with high sensitivity.

AGE // In order to further elucidate different properties of the structural variants, we developed algorithms for identification of the breakpoint at nucleotide level and identification of the mechanism. For this purpose, we developed AGE\cite{21233167}, alignment gap excision, that performs alignment of the sequences at flanking regions of structural variants while considering large deletion and insertion blocks. The conventional global and local sequence alignment algorithms cannot handle the alignment of sequences to breakpoints because the large insertions and deletions cannot be appropriately handled by these tools.

BreakSeq // For identification of mechanism of formation, we developed BreakSeq\cite{20037582}, a pipeline that takes as input the breakpoints at nucleotide resolution and classifies them with respect to formation mechanism. We have recently updated this software significantly as part of Phase 1 of 1000 Genomes \cite{1000-genomes-breakpoints}. We will use the output of BreakSeq pipeline for effect prediction and prioritization of the SVs and CNVs.

Retro-elements // Aiming to better characterize the retrotransposons polymorphisms, a special class of SVs, we built a novel, unified pipeline for retrotransposons discovery using combined evidence from exon-exon junctions, discordant read pairs and read depth\cite{24026178}. We will further polish this pipeline and use it to identify the retrotransposons.

** Variant Prioritization: GENES

*** Preliminary results (GENES)

We have extensive experience in network studies and functional interpretation of coding mutations. Considering diverse gene functions, variants in genes can have a wide spectrum of global effects, ranging from fatal for essential genes to no obvious damaging effect for loss-of-function tolerant genes. The global effect of a coding mutation is largely governed by the diverse biological networks in which the gene participates. 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 position on the top level of regulatory networks \cite{22955619}. Incorporating multiple network and evolutionary properties, we have developed a computation method - NetSNP \cite{23505346} to quantify indispensability of each gene in the genome. The method shows its strong potential for interpretation of variants involved in Mendelian diseases and in complex disorders probed by genome-wide association studies.

While NetSNP identifies and ranks genes, we also plan to develop approaches to quantify variant-specific effects. 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}. Loss-of-function mutations, the most severe form of coding changes have attracted lots of interest in clinical studies. We have published a paper where we have systematically surveyed LoF variants in a cohort of 180 healthy people as part of the Pilot Phase of the 1000 Genomes project \cite{22344438}. Using linear discrimination analysis, we developed a method to distinguish LoF-containing recessive genes from benign LOF-containing genes. In this grant, we will substantially expand this analysis in an effort to understand the impact of LoF variants. Specifically, we propose to develop methods that will (1) provide variant-specific functional impact scores (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.

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}. 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}.

*** Research plan (GENES) : ALOFT + pseudofilter

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. Even LoFs in the same gene cause different phenotypic effect. Another concern about LoF variants are potential calling errors. As shown in \cite{22344438}, LoF variants are prone to calling artifacts. To help filter high-confident LoFs and annotate functional impact of LoFs, we will develop ALoFT to annotate each LOF variants with mismapping, functional, evolutionary and network features. We will quantify the confidence of LoFs using features such as whether it is in highly duplicated regions, number of paralogs and pseudogene and whether it is in 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 variant, as well as truncated sequences. For network features, we quantify how close the gene with LoF to known disease causing genes.

Finally we will develop a machine-learning method to quantify whether LoF will cause benign, recessive or dominant disease-causing effect. We decide to use homozygous LoF observed in healthy individual as benign set and classify known disease-causing LoFs as recessive and dominant sets. We will apply various machine-learning methods to get the best model. We will then evaluate our method with multiple independent dataset, 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 variant-level.

We will develop a computational pipeline, adapted from PseudoPipe, to identify regions of homology between the protein coding genes and the whole genome. We call this pseudofilter. Variants called from these regions will be deprioritized since they might arise due to mismapping of reads derived from a different region of the genome. A two-step process will be used to identify gene homology. First, we will search homology of the gene sequences in the human genome by using BLAT\cite{11932250}, followed by filtering of the alignment results, such as removing the matching sequences whose lengths are shorter than the sequenced read length. Next, we will refine the BLAT results by pairwise alignments. Homology matches from pairwise alignments will be assessed using a sliding window analysis. The length of sliding window should correspond to the read length. For every window, we will test whether the pair of sequences match perfectly with up to two base-pair mismatches. This pipeline will generate a high resolution of homology where mis-mapping may occur between the genes and the rest of the genome as well as unique regions that have no mapping ambiguity issues.

** Variant Prioritization: ncRNAs

*** Preliminary Results (ncRNAs)

We will develop a pipeline called eleVAR – **ele**vating **VA**riants in **R**NA – for variant prioritization in noncoding RNA (ncRNA). To do this, we will leverage our experience analyzing characteristics 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}.

*** Research Plan (ncRNAs)

The eleVAR pipeline will prioritize deleterious rare genetic variants in genomic loci encoding noncoding RNA. The tool will integrate five key categories of features: biochemical interactions and network context, RNA secondary structure, regulatory motifs, evolutionary conservation (GERP score), and expression levels \cite{21152010}. These features will be integrated into a single score for each variant using an entropy based scoring scheme similar to that previously used in our work on noncoding DNA variants\cite{25273974}.

**** More detail on biochemical interactions and network context

RNA interactions with proteins and miRNA are thought to be key to ncRNA function and regulation \cite{22337053}. We will mine protein-RNA interactions from publically available CLIP-Seq data and investigate miRNA-RNA interactions mined from experiments (CLIP/CLASH) or predicted computationally (TargetScan) to create a compendium of biochemical interactions with RNA \cite{25416797, 24297251, 20371350, 23622248, 21909094}. Similar to our previous work on transcription factor binding sites, we will define protein/miRNA footprints on RNA as sensitive to mutation if they are highly enriched for rare genetic variants \cite{24092746, 25273974}. Our preliminary analyses of publicly available CLIP-Seq data indicate that the binding sites of many RNA-binding proteins are comparable with or even more sensitive to mutation than coding sequences. We will use these biochemical interactions to interpret the network context of our variants, using RNA molecules as nodes and RNA-protein and miRNA-RNA interactions as edges. We will prioritize variants that are bound by multiple factors, and those within whole RNAs that are bound by many RNA-binding proteins.

**** More detail on RNA secondary structure

RNA structure is fundamental to the function of most well-studied noncoding RNAs \cite{24895857}. We will predict RNA secondary structures using RNAshapes and compare properties of structured and unstructed regions \cite{16357029}. Our preliminary results indicate that more rigid RNA structures, such as stems, are under higher selection pressure than other RNA regions, and that those variants that incur a larger free energy change of the structures tend to be rarer in human populations. We will define sensitive regions based on folding free energy and folding z-score cutoffs that are enriched for rare genetic variants.

**** More detail on regulatory 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}. We have found that intron-exon junctions, polyadenylation sites, and intron lariat structures are much more sensitive to mutation than other genomic regions, particularly for motif-breaking variants. Similarly, we will investigate motif level features that are important to the functions of classic ncRNA families, e.g. seed regions in miRNAs and complementary regions of U2 and U6 snRNAs \cite{1423631}. Variants that occur in regulatory motif regions will be scored based on the degree to which they break the motif.

**** More detail on entropy-based scoring scheme

To integrate the above features to generate predictive scores for the deleteriousness of variants occurring in ncRNAs, we will use an entropy-based scoring scheme similar to our FunSeq2 tool \cite{FunSeq2}. The intuition behind this scheme is that the more sensitive a given feature is to genetic variation, the lower the probability of observing variants overlapping the feature within healthy individuals in the 1000 genomes project \cite{23128226}. For each feature overlapping a variant, e.g. interaction with a given protein, we will define p_d as the observed frequency of single nucleotide polymorphisms lying within a ncRNA overlaps the feature of interest. The score for that feature, w_d, is then 1-Shannon entropy for the feature (1). Finally, the total score for the variant is the sum of the scores for all overlapping features.

$$w_d=1+ p_d* \log_2 p_d + (1-p_d)*\log_2(1-p_d) \qquad (1)$$

Since some features–e.g. motif-breaking, network degree, and evolutionary conservation–are continuous, we will use the above scheme, except using the observed frequency of genetic variants occurring in regions with scores as or more deleterious than the score of interest. The combined treatment of discrete and continuous features will enable us to integrate new data and data types as they become available.

** Variant Prioritization: TFBS Regions

*** Preliminary Results (TFBS)

**** We have considerable experience annotating non-coding regulatory regions of the genome

Our proposed work is based on our experience in non-coding annotation. We have made a number of contributions in the analysis of the noncoding genome, as part of our extensive 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of a method called PeakSeq to define the binding peaks of TFs\cite{19122651}, as well as new machine learning techniques\cite{19015141}. In addition, we have also proposed a probabilistic model, referred to as target identification from profiles (TIP), that identifies a given TF's target genes based on ChIP-seq data\cite{22039215}. Furthermore, we have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers\cite{20126643}, which we have partially validated\cite{22950945}. We have also constructed regulatory networks for human and model organisms based on the ENCODE\cite{22955619} and modENCODE datasets\cite{21430782}, and completed many analyses on them\cite{22125477,21177976,20439753,15145574,14724320,17447836,15372033,19164758, 16455753,22955619,22950945,18077332,24092746,23505346,21811232,2160691,21253555}

**** We have extensive experience in relating annotation to variation & based on this experience have developed the prototype FunSeq pipeline for Somatic Variants

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}. We further showed relations between selection and protein network structure, e.g. hubs vs periphery\cite{18077332,23505346}.
In recent studies\cite{24092746,25273974}, we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. FunSeq identifies sensitive and ultra-sensitive regions, i.e. those annotations under strong selection pressure as determined by human population variation. It links each noncoding mutation to target genes and prioritizes them based on scaled network connectivity (compute the percentile after ordering centralities of all genes in a particular network). It identifies deleterious variants in many non-coding functional elements, including transcription-factor (TF) binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites and detects their disruptiveness of TF binding sites (both loss-of and gain-of function events). It also develops a scoring scheme, taking into account the relative importance of various features, to prioritize mutations. By contrasting patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, FunSeq allows identification of candidate non-coding driver mutations\cite{24092746}. Our method is able to prioritize the known *TERT* promoter driver mutations and scores somatic recurrent mutations higher than non-recurrent ones. In this study, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project, with cancer genomics data. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples.

*** Research Plan (TFBS): Convert the prototype FunSeq non-coding somatic variant pipeline to prioritize germline variants and elaborate it with new features

We plan to convert the current FunSeq prototype from its focus on germline variants to allow the identification of rare variants associated with high functional impact. We will do some simple improvements (i.e. incorporating GERP scores and ultra-conserved regions for identifying conserved regions between species) and some major changes outlined below.

**** Identifying gain-of- and loss-of-function mutations for TF binding sites

Loss-of- and gain-of-function variants are more likely to cause deleterious impact\cite{23512712,24092746,21596777,23348503,23348506,23530248,23887589}. When variants occur in TF binding motifs, the change in position-weight matrix (PWM) can be calculated. Variants altering the PWM scores could potentially either decrease (loss of function) or increase (gain of function) the binding strength of transcription factors. Determining the ancestral allele of the variant is essential to resolve between loss-of-function and gain-of-function since the functional impact of the variant reflects the historical event when the polymorphism was first introduced in the human population.

**** Better Identification of Enhancers

As part of ENCODE enhancer prediction group, we are working on predicting confident sets of enhancers in human. We are currently developing a new machine learning framework that utilizes epigenomic features and transcription of enhancer RNA (eRNA) to predict active enhancers across different tissues. In addition to the enhancers identified in the ENCODE project, we will also include enhancers from the Roadmap Epigenomics project and cis eQTLs from the GTEx project as regulatory regions of the genome. The activity of these regions will be measured using the epigenomic marks H3K4me1 and H3K27ac, while DNA methylation will be considered as an inactivity signal. We will collect all bisulfite sequencing, ChIP-seq and RNA-seq data from the Roadmap Epigenomics Project\cite{20944595}. Then we will identify significant associations between regulatory elements and candidate target genes through computing the correlations of active signals and anti-correlations of inactive signals with gene expression levels across different tissue types.

We will use the regulatory element - target gene pairs to connect the non-coding variants into a variety of networks -- e.g. regulatory network, metabolic pathways, etc. We will examine their network centralities, such as hubs, bottlenecks and hierarchies, as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious\cite{23505346,18077332}.

** Variant Prioritization: based on allelic analysis

After performing functional annotation based prioritization of variants, we will further up-weight those associated with allele-specific activity.

*** Preliminary Results (allelic)

A specific class of regulatory variants is one that is related to allele-specific events. These are cis-regulatory variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE)\cite{20567245,20846943}. We have previously developed a tool, AlleleSeq,\cite{21811232} for the detection of candidate variants associated with ASB and ASE. Using AlleleSeq, we have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project\cite{22955620,22955619,24092746}. Overall, we found that these allelic variants are under differential selection from non-allelic ones\cite{22955619,24092746}. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression\cite{22955619}. Furthermore, we have provided the AlleleSeq tool, lists of detected allelic variants, and the constructed personal diploid genome and transcriptome of NA12878 on\cite{0000003}.

*** Research Plan (allelic): Variant prioritization based on allelic activity

The evident regulatory roles of the allele-specific variants assert that they will be useful for identifying functional variants, however, we are not aware of any variant prioritization scheme incorporating allele-specificity. One of the main challenges appears to be that allelic variants are enriched for rare variants \cite{24037378}. Moreover, previous analyses were primarily variant-specific or focused mainly on a single deeply-sequenced individual, GM12878 \cite{22955620,22955619,24092746}. We will identify allelic variants from a large pool of individuals. For this purpose, we will extend the AlleleSeq pipeline to account for the overdispersion of empirical read distributions observed in ChIP-Seq and RNA-Seq datasets \cite{25223782,20671027,22499706}. Allelic variants (rare and common) identified across hundreds of genomes will be aggregated into ‘allelic genomic elements’. Each element will be assigned an ‘allelicity’ score based on its enrichment with allelic variants. This will allow incorporation of ASE and ASB into the main prioritization scheme: input variants from allelic genomic elements will be up-weighted according to their scores.

** Variant Priortization: creating a coherent pipeline

We are going to make a robust software suite that integrates all the functionalities we have mentioned above. This software suite allows maximum flexibility for users to parameterize and customize for their own research purpose. We will host this suite on a user friendly web server. Researchers are also able to download this software and do analyses on their home computers.

We will build this pipeline on top of Apache Spark, Parquet and noSQL database to better scale with cluster computing and archive superior performance on large inputs. As security is becoming a big concern in genomics research, we will be using https link and crypto algorithms to protect users information. Also, we allow users to clean up their computational footprints. Last, we will distribute a well encapsulated download package that includes all the tools and required data files (preprocessed and compressed) to give the choice to user to work on their local computational resources. To address the portability issue, we will provide a download version that has everything contained in a Docker container.

* References

All references indicated by \cite{PMID}, where PMID is the PubMed ID. In press references indicated by \cite{name-of-reference}.

1000-genomes-breakpoints
Alexej Abyzov, Shantao Li, Daniel Rhee Kim, Marghoob Mohiyuddin, Adrian Stuetz, Nicholas F. Parrish, Xinmeng Jasmine Mu, Wyatt Clark, Ken Chen, Matthew Hurles, Jan O. Korbel, Hugo Y. K. Lam, Charles Lee, Mark B. Gerstein. “Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications, in press.*