Mark Gerstein’s lab have extensive experience in large-scale variant calling and interpretation through being active members of the 1000 Genomes Consortium, especially in the analysis working group and the structural variant (SV) and functional interpretation (FIG) subgroups of the consortium. We have a lot of experience in large-scale structural variant calling \cite{21787423, 21293372, 20981092, 23128226}. We have developed a number of SV calling algorithms, including BreakSeq, which compares raw reads with a breakpoint library (junction mapping) \cite{20037582.}, CNVnator, which measures read depth and estimates copy number variation \cite{21324876}, AGE, which refines local alignment \cite{21233167}, PEMer, which uses paired ends \cite{19236709}. We have also developed approaches for quantifying retroduplicaton variation \cite{24026178}, array based approaches to structural variation \cite{19037015} and a sequencing-based bayesian model \cite{21034510}. Applying some of these methods to skin we were able to detect somatic mosaicism \cite{23160490}. Furthermore, we have intensively studied the distinct features of SVs originated from different mechanisms. This indicates specific creation processes and potentially divergent functional impacts \cite{24092746,26028266}.

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}. 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}). In recent studies\cite{24092746,25273974}, we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. 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). It then 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 to 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, our method is able to prioritize the known TERT promoter driver mutations, 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.

Gerstein lab also used allelic variability to prioritize regions of the genome. That is, we prioritize regions that differ in functional genomic response, for example, allele-specific expression and binding, between the maternal and paternal alleles (Allelic activity). Our variant analysis work includes AlleleSeq \cite{21811232}, a computational pipeline to identify allele-specific events, and AlleleDB, our database connecting single nucleotide variants with allele-specific binding and expression.

We developed Variant Annotation Tool (VAT)\cite{22743228} 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. We annotated variants from 1,092 humans in Phase 1 of the 1000GP25 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.

We also used networks as a framework for integrating a great variety of genomic variation/mutation data across individuals and organisms and studying their impact on biological systems. We have found that functionally significant and highly conserved genes tend to be more central in interaction and regulatory networks (i.e. more connectivity is associated with more constraint) -- but not in metabolic pathways, where the highly central genes have more duplicated copies and are more tolerant to loss-of-function mutations \cite{23505346, 19521505}. Moreover, we examined the impact of adaptive evolution to protein interaction networks, and found that proteins under positive selections tend to locate at network periphery \cite{18077332}.