***The Gerstein lab’s experience with developing machine learning and analytic approaches to identify variant effects***

The Gerstein lab has extensive experience in developing tools and methods to prioritize cancer-associated genomic alterations and their downstream effects. These pipelines can be readily combined to provide multiple lines of evidence for prioritizing variants and elucidating their impacts, and they have already been successfully applied to a number of disease variant datasets. The corresponding software code for each tool is computationally efficient, thereby enabling us to scale them to large patient cohorts.

We have developed **ALoFT**, a tool specifically tailored to annotate and predict the disease-causing potential of LoF events [PMID: 28851873]. Short for “annotation of loss-of-function transcripts”, we have used ALoFT to successfully discriminate between LoF mutations that are deleterious in heterozygous states from those that may cause disease in the homozygous state. We analyzed somatic variants in more than 6,500 cancer exomes and demonstrated that variants predicted to be deleterious by ALoFT are enriched in canonical cancer driver genes.

With respect to localized perturbations, we performed a separate study [PMID 27915290] to demonstrate how localized changes in biomolecular frustration can be used to better understand the differential effects of variants in oncogenes and TSGs (Fig. 1). Specifically, these results shed light on potential GoF variants on the surfaces of oncogenes, and LoF variants within the cores of TSGs.

In addition to coding variants, we have developed a tool to prioritize non-coding variants in cancer called **FunSeq** [PMID: 24092746]. In brief, FunSeq prioritizes variants based on network connectivity and their disruptiveness (e.g., by finding motif breakers), by identifying deleterious variants in many non-coding functional elements (including transcription factor binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites). In our published work using FunSeq, we integrated large-scale data from various resources, including ENCODE and the 1000 Genomes Project, with cancer genomics data. By comparing patterns of inherited polymorphisms from 1,092 humans with somatic variants, FunSeq identified candidate non-coding cancer driver mutations.

Fig. 1: Prioritizing the effect of SNVs based on changes in localized perturbations (as measured by frustration).

We have developed statistical methods for the analysis of non-coding regulatory regions. **LARVA** (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichment in non-coding elements by comparing observed mutation counts with expected counts under a whole-genome background mutation model [PMID 26304545]. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can be targeted to coding regions to prioritize genes. We used this tool in a pan-cancer analysis of variants in 760 cancer whole genomes, spanning a number of cancer data portals and published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and non-coding cancer drivers, including the TERT and TP53 promoter.

More recently, we developed and applied SVFX, a tool based on a machine learning framework that quantifies the pathogenicity of structural variants [PMID: 33168059].

***The Gerstein lab’s experience with analyzing whole-genome datasets from cancer cohorts***

We have played key roles in TCGA investigations into prostate [PMID: 26544944] and kidney [PMID: 28780132] cancers. As part of the driver discovery subgroup in PCAWG, we participated in a comprehensive variant prioritization exercise to generate a catalog of driver elements in many cancer cohorts. Furthermore, we have led the PCAWG group to investigate the impact of non-coding mutations on cancer development, progression, and prognosis. As part of this effort, we ran our FunSeq pipeline on each variant (~30 million total somatic mutations among 39 cancer subtypes) in PCAWG. In addition to identifying canonical driver mutations, we identified many high-impact mutations that can potentially influence cancer progression.

 We mined the comprehensive variant dataset from the ICGC/TCGA PCAWG project to demonstrate identify a genes and variants that operate as medium-impact putative passengers [PMID: 32084333].