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. The software code for each tool is computationally efficient, thereby enabling them to be scaled to large patient cohorts.

These tools include ALoFT, short for “annotation of loss-of-function (LoF) transcripts”, which is used to predict the disease-causing potential of LoF events throughout the genome (Balasubramanian et all, 2017). This approach successfully discriminates between LoF mutations that are deleterious in heterozygous states from those that may cause disease in the homozygous state. For example, ALoFT identified from across 6,500 cancer exomes that deleterious somatic variants were enriched in canonical cancer driver genes. In another approach(Kumar et al, 2016), localized perturbations were investigated in order to demonstrate how localized structural changes in physically-defined measures of frustration can be used to better understand the differential effects of variants in oncogenes and tumor suppressor genes (TSGs) (See Figure). These results shed light on potential gain-of-function variants on the surfaces of oncogenes, and LoF variants within the cores of TSGs. More recently, SVFX was developed, which is a software tool that uses a machine learning framework to quantify the pathogenicity of structural variants(Kumar et al, 2020).

We have also developed FunSeq (Khurana et al, 2013, a tool to prioritize non-coding variants based on network connectivity and their disruptiveness within these networks (e.g., such as by finding motif breakers which in turn result in changes to regulatory network links). FunSeq has been used to identify deleterious variants in many non- coding functional elements (including TF binding sites, enhancer elements, and regions of accessible chromatin). This software tool has been used to integrate large-scale data from various resources, including ENCODE and the 1000 Genomes Project, with cancer genomics data. For example, by comparing patterns of inherited polymorphisms from 1,092 humans with somatic variants, FunSeq identified candidate non-coding cancer driver mutations. A further method to analyze non-coding regulatory regions, LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutational enrichment in non-coding elements by comparing observed mutation counts with expected counts under a whole-genome background mutation model (Lochovsky et al, 2015). LARVA includes corrections for biases in mutation rate attributed to DNA replication timing. In a pan-cancer analysis of variants in 760 cancer whole genomes (spanning a number of cancer data portals) LARVA also replicated established coding and non-coding cancer drivers, including the TERT and TP53 promoters.



Figure 4: Prioritizing the effect of SNVs based on changes in localized perturbations (as measured by frustration).

The Gerstein has played key roles in TCGA investigations into PCa (Cancer Genome Atlas Research Network, 2015) as well as kidney cancer (Lee et al, 2017), and is also part of the PCAWG driver discovery subgroup to generate a catalog of cancer driver elements. Furthermore, we have led the PCAWG group to investigate the impact of non-coding mutations on cancer development, progression, and prognosis using the FunSeq pipeline on each variant (~30 million total somatic mutations among 39 cancer subtypes) in PCAWG. In addition to identifying canonical driver mutations, this approach identified many high-impact mutations that can potentially influence cancer progression. Finally, mining the comprehensive variant dataset from the ICGC/TCGA PCAWG project identified genes and variants that operate as medium-impact putative passengers (Kumar et al, 2020).

**References**

Balasubramanian S, Fu Y, Pawashe M, McGillivray P, Jin M, Liu J, Karczewski KJ, MacArthur DG and Gerstein M. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat Commun*. 2017;8:382.

Kumar S, Clarke D and Gerstein M. Localized structural frustration for evaluating the impact of sequence variants. *Nucleic Acids Res*. 2016;44:10062-10073.

Kumar S, Harmanci A, Vytheeswaran J and Gerstein MB. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol*. 2020;21:274.

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GRS, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, Genomes Project C, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C and Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*. 2013;342:1235587.

Lochovsky L, Zhang J, Fu Y, Khurana E and Gerstein M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res*. 2015;43:8123-34.

Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*. 2015;163:1011-25.

Lee BH. Commentary on: "Comprehensive molecular characterization of papillary renal-cell carcinoma." Cancer Genome Atlas Research Network.: N Engl J Med. 2016 Jan 14;374(2):135-45. *Urol Oncol*. 2017;35:578-579.

Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, Harmanci A, Martinez-Fundichely A, Chan CWY, Nielsen MM, Lochovsky L, Zhang Y, Li X, Lou S, Pedersen JS, Herrmann C, Getz G, Khurana E and Gerstein MB. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell*. 2020;180:915-927 e16.