SWOG ITSC Pilot Award 2015:

Analysis Plan:

Identification of variants: As part of the International Pan-Cancer Analysis of Whole Genomes (PCAWG) Initiative, the Gerstein Lab is involved in improving and standardizing variant calling methods, developing new methods for SV calling, identifying noncoding drivers and network and pathway analysis. Tools developed for this initiative will be applied to the current project.
Systematic annotation of the variants: We will annotate variants in the coding and non-coding regions using FunSeq2 [1]. Noncoding annotations derived primarily from ENCODE will include transcription-factor binding sites, DNA-hypersensitive sites, chromatin marks by histone binding, predicted enhancer regions, miRNA and pseudogenes [2]. **Identification of c*andidate coding and noncoding drivers***: We developed FunSeq to prioritize both coding and noncoding variants that could be potential drivers [3]. Briefly, the tool identifies potential regions of high functional impact in noncoding regions by understanding patterns of natural variation in human genomes and comparing these patterns in disease cases. We identified regions in the genome under purifying selection that are enriched for rare alleles using variation data from 1092 individuals (Phase1 of 1000 genomes project [4]). Such regions that we dubbed sensitive and ultrasensitive regions highlight regions that are under strong constraint. Mutations identified in such functionally important regions will be considered potential cancer drivers. We have used this method to successfully identify potential noncoding driver mutations in prostate cancer genomes [3]. We have now enhanced this pipeline to include expanded enhancer predictions, gain and loss of function motif analyses for TF-binding and identification of genes that may be affected by such regulatory variants [1].

References

1. Fu Y, Liu Z, Lou S, Bedford J, Mu X, Yip KY, Khurana E and **Gerstein M**. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*. 2014;15:480.
2. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M and **Gerstein M**. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*. 2012;13:R48.
3. 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 GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, 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.
4. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT and McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56-65.