The majority of advanced metastatic prostate cancer (PCa) cases develop resistance to androgen deprivation therapy (ADT) over time, thereby decreasing the median survival of patients to approximately 3 years (Tomlins, Rhodes et al. 2005). Increased efforts have focused on understanding the complexity of PCa genomics. Studies have revealed the impact gene regulation of the androgen receptor (AR), one of the key factors implicated in PCa pathogenesis(Scher and Sawyers 2005, Takeda, Spisak et al. 2018, Morova, McNeill et al. 2020). However, the mechanisms involved in the response to therapy against these regulatory regions and their regulation by AR remain largely unknown. Gerstein lab and collaborators have the experience to uncover gene regulatory alterations that play a critical role in PCa disease progression due to incomplete knowledge of the non-coding genomic landscape of this malignancy.

We have experience in the prioritization of non-coding variants. We developed FunSeq (Khurana, Fu et al. 2013) and FunSeq2 (Fu, Liu et al. 2014) to prioritize variants. We have developed statistical methods for the analysis of non-coding regulatory regions, such as enhancers. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) (Lochovsky, Zhang et al. 2015) and MOAT (Mutations Overburdening Annotations Tool), (PMID: 26304545). ENCODE Cancer-related encyclopedia companion (ENCODEC) also provides background mutation normalization using epigenetics marks for mutation rate analysis (Zhang, Lee et al. 2020).

We have experience in large-scale whole-exome and whole-genome sequencing studies to reveal the functional importance and contribution of somatic mutations (sM) to the unique clinical course of each cancer in cancers (Balasubramanian, Fu et al. 2017, Zhang, Lee et al. 2020) We have developed several tools for analyzing genomic variants from multiple aspects. We built an agnostic machine-learning-based workflow, called SVFX, to assign a “pathogenicity score” to somatic and germline SVs in various cancer types (Kumar, Harmanci et al. 2020). In addition, we developed a method that identifies drivers and quantifies tumor growth that was based solely on the VAF spectrum for an individual sample (Salichos, Meyerson et al. 2020).

We have the experience for network analysis for cancer-associated cell types, we put regulators into hierarchies and measure their network change (rewiring) during oncogenesis (Lou, Li et al. 2020, Zhang, Lee et al. 2020).

References

Balasubramanian, S., et al. (2017). "Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes." Nat Commun **8**(1): 382.

Variants predicted to result in the loss of function of human genes have attracted interest because of their clinical impact and surprising prevalence in healthy individuals. Here, we present ALoFT (annotation of loss-of-function transcripts), a method to annotate and predict the disease-causing potential of loss-of-function variants. Using data from Mendelian disease-gene discovery projects, we show that ALoFT can distinguish between loss-of-function variants that are deleterious as heterozygotes and those causing disease only in the homozygous state. Investigation of variants discovered in healthy populations suggests that each individual carries at least two heterozygous premature stop alleles that could potentially lead to disease if present as homozygotes. When applied to de novo putative loss-of-function variants in autism-affected families, ALoFT distinguishes between deleterious variants in patients and benign variants in unaffected siblings. Finally, analysis of somatic variants in >6500 cancer exomes shows that putative loss-of-function variants predicted to be deleterious by ALoFT are enriched in known driver genes.Variants causing loss of function (LoF) of human genes have clinical implications. Here, the authors present a method to predict disease-causing potential of LoF variants, ALoFT (annotation of Loss-of-Function Transcripts) and show its application to interpreting LoF variants in different contexts.

Fu, Y., et al. (2014). "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer." Genome Biol **15**(10): 480.

Identification of noncoding drivers from thousands of somatic alterations in a typical tumor is a difficult and unsolved problem. We report a computational framework, FunSeq2, to annotate and prioritize these mutations. The framework combines an adjustable data context integrating large-scale genomics and cancer resources with a streamlined variant-prioritization pipeline. The pipeline has a weighted scoring system combining: inter- and intra-species conservation;loss- and gain-of-function events for transcription-factor binding; enhancer-gene linkages and network centrality; and per-element recurrence across samples. We further highlight putative drivers with information specific to a particular sample, such as differential expression. FunSeq2 is available from funseq2.gersteinlab.org.

Khurana, E., et al. (2013). "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science **342**(6154): 1235587.

Interpreting variants, especially noncoding ones, in the increasing number of personal genomes is challenging. We used patterns of polymorphisms in functionally annotated regions in 1092 humans to identify deleterious variants; then we experimentally validated candidates. We analyzed both coding and noncoding regions, with the former corroborating the latter. We found regions particularly sensitive to mutations ("ultrasensitive") and variants that are disruptive because of mechanistic effects on transcription-factor binding (that is, "motif-breakers"). We also found variants in regions with higher network centrality tend to be deleterious. Insertions and deletions followed a similar pattern to single-nucleotide variants, with some notable exceptions (e.g., certain deletions and enhancers). On the basis of these patterns, we developed a computational tool (FunSeq), whose application to ~90 cancer genomes reveals nearly a hundred candidate noncoding drivers.

Kumar, S., et al. (2020). "SVFX: a machine learning framework to quantify the pathogenicity of structural variants." Genome Biol **21**(1): 274.

There is a lack of approaches for identifying pathogenic genomic structural variants (SVs) although they play a crucial role in many diseases. We present a mechanism-agnostic machine learning-based workflow, called SVFX, to assign pathogenicity scores to somatic and germline SVs. In particular, we generate somatic and germline training models, which include genomic, epigenomic, and conservation-based features, for SV call sets in diseased and healthy individuals. We then apply SVFX to SVs in cancer and other diseases; SVFX achieves high accuracy in identifying pathogenic SVs. Predicted pathogenic SVs in cancer cohorts are enriched among known cancer genes and many cancer-related pathways.

Lochovsky, L., et al. (2015). "LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations." Nucleic Acids Res **43**(17): 8123-8134.

In cancer research, background models for mutation rates have been extensively calibrated in coding regions, leading to the identification of many driver genes, recurrently mutated more than expected. Noncoding regions are also associated with disease; however, background models for them have not been investigated in as much detail. This is partially due to limited noncoding functional annotation. Also, great mutation heterogeneity and potential correlations between neighboring sites give rise to substantial overdispersion in mutation count, resulting in problematic background rate estimation. Here, we address these issues with a new computational framework called LARVA. It integrates variants with a comprehensive set of noncoding functional elements, modeling the mutation counts of the elements with a beta-binomial distribution to handle overdispersion. LARVA, moreover, uses regional genomic features such as replication timing to better estimate local mutation rates and mutational hotspots. We demonstrate LARVA's effectiveness on 760 whole-genome tumor sequences, showing that it identifies well-known noncoding drivers, such as mutations in the TERT promoter. Furthermore, LARVA highlights several novel highly mutated regulatory sites that could potentially be noncoding drivers. We make LARVA available as a software tool and release our highly mutated annotations as an online resource (larva.gersteinlab.org).

Lou, S., et al. (2020). "TopicNet: a framework for measuring transcriptional regulatory network change." Bioinformatics **36**(Suppl\_1): i474-i481.

MOTIVATION: Recently, many chromatin immunoprecipitation sequencing experiments have been carried out for a diverse group of transcription factors (TFs) in many different types of human cells. These experiments manifest large-scale and dynamic changes in regulatory network connectivity (i.e. network 'rewiring'), highlighting the different regulatory programs operating in disparate cellular states. However, due to the dense and noisy nature of current regulatory networks, directly comparing the gains and losses of targets of key TFs across cell states is often not informative. Thus, here, we seek an abstracted, low-dimensional representation to understand the main features of network change. RESULTS: We propose a method called TopicNet that applies latent Dirichlet allocation to extract functional topics for a collection of genes regulated by a given TF. We then define a rewiring score to quantify regulatory-network changes in terms of the topic changes for this TF. Using this framework, we can pinpoint particular TFs that change greatly in network connectivity between different cellular states (such as observed in oncogenesis). Also, incorporating gene expression data, we define a topic activity score that measures the degree to which a given topic is active in a particular cellular state. And we show how activity differences can indicate differential survival in various cancers. AVAILABILITY AND IMPLEMENTATION: The TopicNet framework and related analysis were implemented using R and all codes are available at <https://github.com/gersteinlab/topicnet>. SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.

Morova, T., et al. (2020). "Androgen receptor-binding sites are highly mutated in prostate cancer." Nat Commun **11**(1): 832.

Androgen receptor (AR) signalling is essential in nearly all prostate cancers. Any alterations to AR-mediated transcription can have a profound effect on carcinogenesis and tumor growth. While mutations of the AR protein have been extensively studied, little is known about those somatic mutations that occur at the non-coding regions where AR binds DNA. Using clinical whole genome sequencing, we show that AR binding sites have a dramatically increased rate of mutations that is greater than any other transcription factor and specific to only prostate cancer. Demonstrating this may be common to lineage-specific transcription factors, estrogen receptor binding sites were also found to have elevated rate of mutations in breast cancer. We provide evidence that these mutations at AR binding sites, and likely other related transcription factors, are caused by faulty repair of abasic sites. Overall, this work demonstrates that non-coding AR binding sites are frequently mutated in prostate cancer and can impact enhancer activity.

Salichos, L., et al. (2020). "Estimating growth patterns and driver effects in tumor evolution from individual samples." Nat Commun **11**(1): 732.

Tumors accumulate thousands of mutations, and sequencing them has given rise to methods for finding cancer drivers via mutational recurrence. However, these methods require large cohorts and underperform for low recurrence. Recently, ultra-deep sequencing has enabled accurate measurement of VAFs (variant-allele frequencies) for mutations, allowing the determination of evolutionary trajectories. Here, based solely on the VAF spectrum for an individual sample, we report on a method that identifies drivers and quantifies tumor growth. Drivers introduce perturbations into the spectrum, and our method uses the frequency of hitchhiking mutations preceding a driver to measure this. As validation, we use simulation models and 993 tumors from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium with previously identified drivers. Then we apply our method to an ultra-deep sequenced acute myeloid leukemia (AML) tumor and identify known cancer genes and additional driver candidates. In summary, our framework presents opportunities for personalized driver diagnosis using sequencing data from a single individual.

Scher, H. I. and C. L. Sawyers (2005). "Biology of progressive, castration-resistant prostate cancer: directed therapies targeting the androgen-receptor signaling axis." J Clin Oncol **23**(32): 8253-8261.

Prostate cancers that are progressing on medical and surgical therapies designed to ablate the action of androgens continue to express androgen receptor (AR) and to depend on signaling through the receptor for growth. A more clinically relevant classification of castration-resistant disease focuses on the mechanisms of receptor activation, which include (1) changes in the level of ligand(s) in tumor tissue; (2) increased levels of the protein due to gene amplification or altered mRNA expression; (3) activating mutations in the receptor that affect structure and function; (4) changes in coregulatory molecules including coactivators and corepressors; and (5) factors that lead to activation of the receptor independent of the level of ligand or receptor allowing kinase cross talk. From an AR perspective, the term "hormone refractory" is inappropriate. On the basis of this schema, we discuss strategies that are focused on the AR either directly or indirectly, as single agents or in combination, that are in clinical development.

Takeda, D. Y., et al. (2018). "A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer." Cell **174**(2): 422-432 e413.

Increased androgen receptor (AR) activity drives therapeutic resistance in advanced prostate cancer. The most common resistance mechanism is amplification of this locus presumably targeting the AR gene. Here, we identify and characterize a somatically acquired AR enhancer located 650 kb centromeric to the AR. Systematic perturbation of this enhancer using genome editing decreased proliferation by suppressing AR levels. Insertion of an additional copy of this region sufficed to increase proliferation under low androgen conditions and to decrease sensitivity to enzalutamide. Epigenetic data generated in localized prostate tumors and benign specimens support the notion that this region is a developmental enhancer. Collectively, these observations underscore the importance of epigenomic profiling in primary specimens and the value of deploying genome editing to functionally characterize noncoding elements. More broadly, this work identifies a therapeutic vulnerability for targeting the AR and emphasizes the importance of regulatory elements as highly recurrent oncogenic drivers.

Tomlins, S. A., et al. (2005). "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer." Science **310**(5748): 644-648.

Recurrent chromosomal rearrangements have not been well characterized in common carcinomas. We used a bioinformatics approach to discover candidate oncogenic chromosomal aberrations on the basis of outlier gene expression. Two ETS transcription factors, ERG and ETV1, were identified as outliers in prostate cancer. We identified recurrent gene fusions of the 5' untranslated region of TMPRSS2 to ERG or ETV1 in prostate cancer tissues with outlier expression. By using fluorescence in situ hybridization, we demonstrated that 23 of 29 prostate cancer samples harbor rearrangements in ERG or ETV1. Cell line experiments suggest that the androgen-responsive promoter elements of TMPRSS2 mediate the overexpression of ETS family members in prostate cancer. These results have implications in the development of carcinomas and the molecular diagnosis and treatment of prostate cancer.

Zhang, J., et al. (2020). "An integrative ENCODE resource for cancer genomics." Nat Commun **11**(1): 3696.

ENCODE comprises thousands of functional genomics datasets, and the encyclopedia covers hundreds of cell types, providing a universal annotation for genome interpretation. However, for particular applications, it may be advantageous to use a customized annotation. Here, we develop such a custom annotation by leveraging advanced assays, such as eCLIP, Hi-C, and whole-genome STARR-seq on a number of data-rich ENCODE cell types. A key aspect of this annotation is comprehensive and experimentally derived networks of both transcription factors and RNA-binding proteins (TFs and RBPs). Cancer, a disease of system-wide dysregulation, is an ideal application for such a network-based annotation. Specifically, for cancer-associated cell types, we put regulators into hierarchies and measure their network change (rewiring) during oncogenesis. We also extensively survey TF-RBP crosstalk, highlighting how SUB1, a previously uncharacterized RBP, drives aberrant tumor expression and amplifies the effect of MYC, a well-known oncogenic TF. Furthermore, we show how our annotation allows us to place oncogenic transformations in the context of a broad cell space; here, many normal-to-tumor transitions move towards a stem-like state, while oncogene knockdowns show an opposing trend. Finally, we organize the resource into a coherent workflow to prioritize key elements and variants, in addition to regulators. We showcase the application of this prioritization to somatic burdening, cancer differential expression and GWAS. Targeted validations of the prioritized regulators, elements and variants using siRNA knockdowns, CRISPR-based editing, and luciferase assays demonstrate the value of the ENCODE resource.