**psychEncode experience related to QTLs**

We have extensive experience in identifying QTLs of various types, as well as relating these QTL types to one another. We leveraged uniformly-processed data from 1866 individuals {Wang, 2018 #32}to identify QTLs originating from ~1.3 million, 7,976 and 1,672 SNPs, which were associated with eQTLs, cQTLs, and cell fraction QTLs (fQTLs), respectively. In order to identify cis-eQTLs, we closely adhered to the protocols used by GTEx, as well as the standardized approach established by ENCODE to uniformly process RNA-seq data. Furthermore, we benchmarked our results with direct comparisons to available data files in the GTEx portal (gtexportal.org){Consortium, 2015 #31}. We followed existing protocols to ensure compatibility and comparison between our results and those previously published. Using these methods in PsychENCODE I {Wang, 2018 #32}, we identified a set of eQTLs several times as large as those in previous studies (targeting a saturating proportion of protein-coding genes). We found ~2.5 million eQTLs comprising ~238,000 linkage-disequilibrium–independent SNPs. The lists of significant eQTLs are available at resource.psychencode.org. Our set of the eQTLs were found to be consistent with those reported by GTEx. We measured the similarity between PsychENCODE brain dorsolateral prefrontal cortex (DLPFC) eQTLs and GTEx eQTLs of other tissues using pi\_1 values and SNP-eGene overlap rates. We found that both pi\_1 values and SNP-eGene overlap rates are higher for brain DLPFC than for the other tissues. In a separate analysis, we intersected our eQTLs with other QTL types (such as cQTLs and fQTLs). We found that eQTL-cQTL intersections often suggested that an expression-modulating function of an eQTL was derived from chromatin changes (an example which we discussed in our study was that related to mTOR).

We also have identified and studied trans-eQTLs. We matched RNA-seq and genotype data to generate trans-eQTLs using QTLtools, and comprehensively controlled for false-positives. This led to the identification of ~562K trans-eQTLs and found that more than 80% of the trans-eGenes overlapped with cis-eGenes. We then explored the possible mechanisms for trans-eQTL associations. We found that some cis-eGenes and trans-eGenes had a TF-target gene relationship. Trans-eQTL variants and trans-genes also showed higher interaction frequencies than randomly-selected pairs. These findings indicated that trans-eQTL variants interacted with genes via inter-chromosomal interactions. We also found that if one trans-eQTL variant is associated with multiple trans-genes, those trans-genes show higher co-expression patterns than randomly-selected genes.

Finally, we used these QTL types (including fQTLs) as a way to build links within a bulk gene regulatory network. Associations between genomic variants and *intermediary phenotypes* – such as expression eQTLs, cQTLs), and fQTLs –define regulatory relationships within this network. The large number of samples used to calculate these QTLs (and thereby the large number of links within the regulatory network), has provided a substantially larger dataset of brain-based QTLs than previously available. We also evaluated the intersections between these various QTL types (i.e., the frequency with which the SNPs between different QTL types were shared), thereby allowing us to select “multi-QTLs”, which are highly penetrant in that they influence multiple intermediate phenotypes.

**Variant effects at multiple levels**

We have experience in estimating the effects of noncoding variants using data integration based on 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. Using data from multiple sources, we have developed a weighted scoring system to prioritize variants based on their predicted phenotypic effect. We also developed RADAR which extends FunSeq2 {Fu, 2014 #34} by combining RNA regulome and tissue-specific information. We have also built the GRAM pipeline {Lou, 2019 #38}, which is a generalized framework to predict the cellular expression modulating effect of a noncoding variant by incorporating TF disruptive information, histone modification, and cell-specific expression and regulatory network information. GRAM can be extensively applied to fine map causal variants within an LD associated region.

**Prior experience related to characterizing SVs and their breakpoints**

We have extensively studied the characteristics of SVs originating from different mechanisms. This indicates specific creation processes and potentially divergent functional impacts {Khurana, 2013 #39}{Abyzov, 2015 #1}. For example, we observed that NAHR-mediated SVs are associated with open chromatin environments and the activation of enhancers. Our analyses have also shown that micro-insertions flanking breakpoints of NHR-mediated SVs are templated from late replicating genomic regions.Hence, we can gain insights into the functional impacts of different types of SVs types by knowing how they were formed. Hence, we can gain insights into the functional impacts of different types of SVs types by knowing how they were formed. We also performed SV mechanism annotations for the 1000 Genomes Phase 3 deletions using BreakSeq {Lam, 2010 #40}; and categorized 29,774 deletions into NAHR, NHR, TEI and VNTR-based mechanisms. Among these, NHR is the most prevalent mechanism (~73% of all categorized deletions) {Sudmant, 2015 #14}. Additionally, we have also quantified the constraint induced by SVs on various genomic elements {Sudmant, 2015 #14}. Our analysis highlighted that SVs affecting genic regions (coding, UTRs, and intronic) are highly depleted. Moreover, the observed depletions are relatively higher among those SVs that are partially overlapping with genic regions compared to those SVs that are completely engulfing them. Similarly, SVs are also depleted among regulatory elements including enhancers and TF binding sites. In contrast, we found SVs to be enriched among pseudogenes consistent with their formation mechanism, which involves either duplication or retrotransposition.

**Our supervised machine learning framework to assign pathogenicity to SVs**

Recently, we have developed a supervised machine learning based framework to prioritize somatic and germline SVs. The underlying hypothesis of our approach is that various genomic and epigenomic features of disease SVs are very distinct from benign SVs observed among healthy populations. Furthermore, such differences can be sensitively identified in appropriate tissue specific contexts. Thus, we have developed tissue specific models that quantify pathogenicity score by comparing genomic and tissue-specific epigenomic features of a given SV to known benign SVs. To build our somatic and germline machine learning models, we leveraged SVs from the Pan-Cancer Analysis of the Whole Genomes (PCAWG) Project {Consortium, 2020 #42}, Genome Sequencing Program (GSP) and the 1000 Genomes (1KG) Project {Sudmant, 2015 #14}. We employed tissue-specific epigenomic data from Epigenome Roadmap, various genomic element annotations, and cross-species conservation metrics to generate our feature sets. Overall, our approach achieved high accuracy in identifying pathogenic somatic deletions (mean AUC of 0.865) and duplications (mean AUC 0.835) across multiple cancer types. Similarly, our germline models showed high accuracy in identifying pathogenic germline SVs in cancer (mean AUC 0.8) as well as cardiovascular disease (mean AUC 0.79).

Moreover, we also evaluated the contribution of various features toward performance of our somatic and germline models. As expected, we found SV length and overlap with conservation related features had the most significant contributors to the predictive performance for these models. Additionally, various noncoding and epigenomic features, including overlap with 3' and 5' UTR, sensitive regions and H3K4me3 signals had significant contribution toward the predictability of our somatic deletion models suggesting an essential influence of SVs on cis-regulatory elements. Similarly, the predictive performance of somatic duplication models primarily depends on overlap with known heterochromatin annotation, UTRs, and sensitive regions. In our germline model, along with various regulatory region features, we also observed strong contribution of 3D-genome annotation toward predictability of our models.

**Prior experience in conducting workshop and hackathon at Yale Center for Biomedical Data Science workshop**

The Center for Biomedical Data Science (CBDS) at Yale has extensive experience in designing, organizing, and implementing university-wide workshops to train physicians and scientists around Yale in putting machine learning methods to direct practice. During the Summer of 2019, CBDS offered a one-week course titled “Introduction to Informatics & Data Science in the Clinical Health and Biomedical Context ''. Yale postdocs and research scientists from the Gerstein lab designed components of and co-taught this course (along with associated hands-on-lab components) on machine and statistical learning. Enrollment comprised a combination of ~45 junior faculty members, physicians, and postdocs in the life sciences, and topics that were covered included random forests, support vector machines, neural networks, linear regression, and Bayesian-based methods. Student reviews of the course suggested that this workshop was highly valued by participants, thereby prompting CBDS to offer the same workshop/course again the following Summer.

In addition, CBDS has also offered a course titled “Machine Learning for Single-cell Analysis”. This course was offered to a similar audience to that described above, and it was taught in October 2019. The purpose of this three-day workshop was to clarify many of the complex aspects behind real-world single-cell analytical techniques. Course participants gained practical skills for working with single cell datasets and implemented machine learning foundations behind each method. Students also receive an introduction to new trends in deep learning techniques as they are implemented in the context of single cell analysis. As with the machine learning course detailed above, course participants heard lectures as well as worked on hands-on labs during each day of the course.