Previous experiences related to protein 3D-structure to interpret and prioritize SNPs

Using models of protein conformational change, we developed a comprehensive computational framework that incorporates protein structure and dynamics for predicting allosteric residues both on the surface and in the interior. Computational efficiency has been a priority in the development of this framework, thereby enabling high- throughput analysis for large protein datasets. To identify likely ligand binding sites, we use a modified version of the binding leverage method introduced by Mitternacht and Berezovsky {\cite 21935347}. We score surface sites based on the degree to which deformations at a given site couple to modeled conformational changes (such sites are then predicted to be allosteric). Interior allosteric residues are predicted on the basis of whether or not they lie within crucial communication pathways between subregions within proteins. Using these approaches, we found that these predicted allosteric residues tend to be conserved over diverse evolutionary time scales. In particular, we used ExAC variant data to demonstrate the relative conservation of these sites across human populations. In addition, we leveraged HGMD and ClinVar variant data to identify proteins with predicted allosteric residues that coincide with disease associated SNVs. Notably, a number of the disease-associated variants which map to our predicted allosteric sites had previously been poorly understood (in that alternative mechanisms for pathogenicity had not been readily apparent).

Rare variants present challenges in evaluating potential deleteriousness using simple phenotype-genotype associations. We have expertise in using metrics of localized structural frustration perturbations as a means of measuring the severity of impacts conferred by both rare and common variants in protein structures \cite{ 27915290}. Our approach, as applied to the PDB, demonstrates that localized frustration uncovers a number of intuitive and biologically interesting results. Firstly, disease-associated variants give rise to more extreme changes in localized frustration than do non-disease related variants, and rare SNVs tend to disrupt local interactions to a larger extent than common variants.

Prior experience with structural variations

We have extensive experience in identifying as well as interpreting large genomic structural variations as part of the 1000 Genomes Project and the Human Genome Structural Variation (HGSV) Consortium. In particular, we played an active role in the analysis working group and the structural variant and functional interpretation subgroups of the consortium, where the majority of the variant calling tools were developed, deployed and interpreted. We have developed a number of SV calling algorithms, including BreakSeq \cite{ 20037582} which compares raw reads with a breakpoint library (junction mapping, CNVnator \cite{ 21324876} which measures read depth, AGE \cite{} which refines local alignment, and PEMer \cite{19236709} which uses paired end reads. We have also developed array-based approaches and a sequencing-based Bayesian model. Furthermore, we have studied the distinct features of SVs that originate from different mechanisms and showed how creation processes may have potentially divergent functional impacts. We have also played an active role as part of the HGSV consortium to understand the improvement in structural variant calls by applying long-read sequencing platforms such as PacBio, Bionano, Strand-seq and Oxford-Nanopore \cite{30992455}.

Along with SV discovery, we have also developed a supervised machine learning based framework to prioritize somatic and germline SVs (Kumar et al. 2019 bioRxiv). 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. We have trained these models for multiple disease cohorts including cancer, Cardiovascular disease, and Inflammatory Bowel Disease. Our model performed very well for independent datasets in these distinct disease cohorts with high sensitivity and specificity (mean AUC in range of 0.865-0.8).

Previous experience with mutation burdening analysis

We have extensive experience in performing genomic burden analysis in somatic settings. These approaches are highly flexible and can be easily extended in the rare disease context. For instance, we developed a computational framework called LARVA, which integrates variants with a set of non-coding functional elements to model mutation counts of the elements and handle overdispersion \cite{26304545}. This framework incorporates regional genomic features such as replication timing to better estimate local mutation rates and finds mutational hotspots. We have identified well-known non-coding drivers and uncovered new potential non-coding driver regions after applying LARVA to hundreds of whole-genome tumor sequences.

Similarly, we have also performed burdening analysis on protein structure level to identify significantly mutated regions and the corresponding cancer driver genes \cite{31462496}. In this approach, we integrated protein 3D-structure along with protein dynamics information to identify presence of mutational hotspot communities on protein structure. In the somatic setting, these hotspot communities correspond to groups of residues that have correlated motion and are enriched for cancer mutations compared to a background expectation. We will extend this framework in the germline setting by identifying those residue communities that are under high constraint for harboring germline mutations in healthy population. This will generate a comprehensive set of constrained subunits on protein structures, which we can leverage to prioritize missense mutations in the rare disease context.

Previous experience on RNA-seq

We have expertise in data analytics and leading large-scale national projects whose focus is to develop and disseminate novel tools and methods. For instance, we have developed methods for normalization, analysis, and comparison of bulk RNA-seq profiles \cite{21177971,21349863}. Furthermore, we have developed multiple advanced mathematical and machine learning methods covering a variety of key tasks to analyze large-scale snRNA-seq data, including but not limited to preprocessing, imputation, quantification, visualization, clustering, and comparisons. We also have experience with predicting gene expression levels from chromatin features in ENCODE. Using ENCODE data, we integrated RNA-seq data and chromatin features in multiple cell lines, DNase I hypersensitivity site information, and other data types with protein-protein interaction and transcriptional regulatory networks to group TFs into histone-sensitive and histone-insensitive classes. These classification may be used to refine the prediction of gene-regulation targets and effects \cite{22950368}. Finally, as part of the exRNA consortium, we developed the extracellular RNA processing toolkit (exceRpt) pipeline for uniform processing throughout the consortium \cite{30956140}. exceRpt has been used over 80,000 times. exceRpt performs sequential alignment of RNA to contaminants, to human transcriptome and genome sequences, to human repetitive elements, and finally to exogenous sequences.

Data Dissemination

Previous work related to Genomic Privacy

We have developed a formalism for eQTL leakage from gene expression values and SV leakage from signal profiles of functional genomics data {\cite 26828419}. We showed that through various linking attack scenarios we were able to connect individuals to databases of personal information. In the first scenario, we used the inference of the correlated eQTLs with the extreme gene expression values and used them to genotype individuals. We showed that these genotypes yield more than 90% linking accuracy when applied to the gEUVADIS dataset. In the second scenario, we focused on the signal profiles from functional genomics data. The attack involves cross-referencing the individuals in a signal profile dataset, S, against the individuals in a genotype dataset, G. The main objective of the adversary is to link G and S by first predicting the genotypes using signal profiles in S, and then matching the predicted genotypes to the genotypes in G. For any matching individuals in G and S, the name and sensitive information are revealed to the adversary. We also developed tools to deal with the leakage from signal tracks {\cite 29934598}. The most effective way to protect against a linking attack scenario is to ensure that deletion genotypes cannot be inferred from signal profiles. Deletions are a major source of leakage of genetic information from functional genomics signal profiles. We proposed solutions to the signal profiles where we can mask the sensitive information leakage, while providing high utility. Our proposed solution systematically removes the dips in signal profiles as a way to anonymize the profiles against the prediction of deletions. To remove these dips systematically, we used median filtering-based signal processing to locally smooth the signal profile around the deletion. This signal processing technique has been used to remove shot noise in two-dimensional imaging data and one-dimensional audio signals.