Dr. Gerstein’s lab has extensive experience and their involvement in the consortium will offer much needed expertise to bring state-of-the-art methodologies to process, analyze and interpret the data to be collected. Dr. Gerstein developed an interpretable Deep Structured Phenotype Network (DSPN) [30545857], a deep learning model based on conditional Deep Boltzmann Machine architecture with multiple layers. For cancer, he participated in the Pan-Cancer Analysis Working Group (PCAWG)/TCGA [32025007].

Omic integration.For the integration between genetics and multi-omics data, Dr. Gerstein has developed several methods for large-scale human and model organism genome analysis projects (e.g., ENCODE [22955616,22955619], modENCODE [21177976], PsychENCODE [30545857, 26605881], the Extracellular RNA Communication Project  [30956140, 30951667], GENCODE [22951037,30357393], Impact of Genomic Variation on Function (RFA-HG-20-043), Human Genome Structural Variation Consortium, Centers for Mendelian Genomics [35148959], Pan-cancer analysis of whole genomes, Cancer Genome Atlas [32025007], 1000 Genomes [REF], and Developmental Genotype-Tissue Expression (RFA-HD-21-008). Dr. Gerstein has been responsible for projects related to brain disorders [30545856; 30545854; 36323788; https://www.biorxiv.org/content/10.1101/2021.09.07.459322v1] as well as several asthmas [32571363,33059594] and cancer projects [32084333, 32025015, 32024998, 32024824, 32681003, 32728046].

Statistical and machine learning approaches to multi-omics data analysis. Dr. Gerstein’s group has created advanced capability to make use of multi-omics data for patient clustering [36339813], classification [29212468], pathway inference [30956747], outcome prediction [26980320], and to identify genes for complex traits [36216496]. Dr. Gerstein’s team developed an initial single-cell ATAC-seq pipeline and simulator to identify enhancers from single-cell multi-omics data [33471102]. For the PsychENCODE consortium, Dr. Gerstein’s team developed tools [30545857] that combine enhancer, bulk and single-cell RNA-seq, genotype, transcriptome, chromatin, and Hi-C dataset to unravel the underlying mechanism of psychiatric disorders. These tools are extendable to include additional data types such as neuroimaging and can be generalized to other diseases. They also did much work for the integrative analysis of multi-omic datasets from ENCODE [22955619, 25164757,22955616,22955620,25164755] and modENCODE [25164755,21177976] consortia [21926158,22955978].

Transcriptional data. Gerstein’s team has deployed multiple data processing workflows for transcriptional data, including RNA seq [22238592, 21134889, 32584815] and extracellular RNA processing toolkit (exceRpt) pipeline for the exRNA consortium [30956140, 30951667] to perform sequential alignment of RNA. Gerstein’s team developed several tools including: 1) STARRPeaker [33292397] for the uniform processing of STARR-seq data; 2) MrTADFinder [28742097] to identify topologically associating domains from Hi-C data; and 3) tools for enhancer prediction and boundary localization, such as DECODE and Matched-filter [34252960,32737473].

Variant Identification and Prioritization. The following is a summary of methods we have developed for prioritizing germline and somatic variants used for multi-institutional data generation projects (e.g. 1,000 genomes and TCGA) [24092746; 20981092; 32025007]: 1) annotated putative loss-of-function variants  [22344438; 28851873] to infer  the pathogenicity of pLoF variants in Mendelian diseases, autism, and cancer; 2) the FunSeq Variant-prioritization pipeline that identifies disease-causing mutations broadly across many conditions [24092746; 25273974]; 3) extensions of FunSeq for structural variant detection [32727537; 31469829] which can be used in combination with other structural variant pipelines we developed [26432245,33168059, 20037582]; 3) LARVA which identifies statistically significant mutation enrichments in somatic non-coding elements [26304545]; 4) MOAT, for burden analysis [29121169]; 5) NIMBus which successfully identified mutational hotspots that potentially disrupt gene regulatory networks in cancer [33092526]; 6) PCAWG ascertains the molecular functional impact of each variant and shows that in addition to high- and low-impact mutations, there is a group of medium-impact putative passengers predicted to influence gene activity [32084333].

Diagram

Description automatically generated

*Figure 1. FunSeq is specialized to prioritize somatic variants from cancer whole genome sequencing.*

Large-Scale Integration. In their role of facilitating various human genome projects, including the ENCODE, modENCODE, and GENCODE, Dr. Gerstein integrated multiple genomic datasets to construct gene regulatory networks consisting of regulatory factors including transcription factors and micro-RNAs, and their target genes [21177976,25164755, 22955619]. Dr. Gerstein is a major participant in GENCODE with a focus on pseudogenes [25157146, 22951037] and genome annotation [30357393]. As part of the Data Analysis Coordination Center in the IGVF consortium, Gerstein’s team contributes to providing the biomedical community with a catalog of human genetic variants. They also led the PsychENCODE data analysis and used the regulatory network based on Hi-C, QTLs and activity relationships to connect non-coding genome-wide association study loci to potential psychiatric disease genes including schizophrenia, autism, bipolar disorder and Alzheimer's [30545857]. Gerstein’s team previously integrated multiple genomic datasets to construct gene regulatory networks, consisting of various regulatory factors including transcription factors and micro-RNAs and their target genes [22955619,25164757,22125477]. Dr. Gerstein developed many integrative tools within the framework of consortiums mentioned above. For example, his team applied topic modeling to integrate multiple data sources including gene expression [32571363] to identify gene groups that show significant co-expression behaviors and related functions [32657410]. In another example, they constructed a Multinet [23505346] using data from various biological networks (PPI, phosphorylation, signaling, metabolic, genetic and regulatory) to analyze genes via their roles in the individual and combined networks and to assign functional indispensability scores to genes. They also developed different methods for determining network hierarchies, such as HirNet [25880651].

Diagram, schematic

Description automatically generated *Figure 2. Human Transcription-factor–miRNA regulation derived from ENCODE data*