Gerstein Lab experience in developing Genomic Tools.

We have extensive experiences in employing and constructing genomics tools. In one work, we presented exceRpt, the exRNA-processing toolkit of the NIH Extracellular RNA communication Consortium (ERCC), which generates quality control and abundance estimates for RNA biotypes and characterizes mappings to exogenous genomes for phylogenetic tree analyses (30956140). In addition, we developed a two-pass approach for scoring ChIP-Seq relative to controls. Specifically, the first pass identifies putative binding sites and compensates for genomic variation in sequence mapping while the second pass filters binding sites not significantly enriched, computing precise enrichments for sites that are significantly enriched relative to normalized controls (19122651). In another example, we presented MUSIC, a signal processing approach for identification of enriched regions in ChIP-Seq data by filtering the read-depth signal for noise from variation in sequence mapping and using a multiscale decomposition for identifying enriched regions at different length scales (https://doi.org/10.1186/s13059-014-0474-3). The proposed approach performs better in terms of accuracy and reproducibility when compared to other models (https://doi.org/10.1186/s13059-014-0474-3)

Furthermore, we have broad interests and experiences in next-generation sequencing. In one work, we developed IQSeq (Isoform Quantification in next-generation Sequencing) to reconstruct the abundances of isoforms of a gene for improved precision of measuring transcription in RNA-Seq experiments (22238592). IQSeq not only allows users to optimally model the integration of various sequencing-technologies in a cost-effective way with features relevant to a particular experimental design, but also supports a generalized statistical partial-sample-generation function with “plugin-able” function for specific sequencing technologies (22238592). In another work, we presented a computational pipeline (AlleleSeq) that constructs a diploid personal genome sequence using genomic variants to identify allele-specific events that have significant differences between the maternally and paternally alleles (21811232). In another work, we developed the Local Event-based analysis of alternative splicing using RNA-Seq pipeline by leveraging information of local splicing events to identify unambiguous alternative splicing and quantify the abundance of these events, enabling a robust discovery of differential alternative splicing (https://doi.org/10.1101/841494). In addition, we addressed the challenge of non-uniform, and over dispersed coverage which is further confounded by RNA secondary structure and thermodynamic stability in STARR-seq technology by developing STARRPeaker for uniform processing of the data (33292397).

Gerstein Lab experience in single-cell analysis. We developed several tools for analyzing single cell datasets from multiple aspects. To simulate single cell ATAC-seq (scATAC-seq) experiments. We developed SCAN-ATAC-Sim to simulate scATAC-seq data by down-sampling bulk ATAC-seq data of representative cell lines or tissues. SCAN-ATAC-Sim uses a consistent but tunable signal-to-noise ratio across cell types to simulate bulk experiments with various levels of background noise in a scATAC-seq simulation. This tool is useful for evaluating the efficacy of single-cell data analysis techniques by simulating scATAC-seq data while adhering to various biological constraints. The multicore parallelism acceleration permits the simulation of millions of cells in less than an hour. Moreover, this tool is highly scalable and offers a space-time tradeoff to match growth rate of cell numbers sequenced by scATAC-seq (33471102).

We also developed a collection of tools to analyze the single cell RNA-seq (scRNA-seq) data. For single cells clustering, we developed Forest Fire Clustering, which is a scalable and interpretable method to cluster single cells and discover cell types using scRNA-seq data. Forest Fire Clustering makes minimal prior assumptions and calculates a non-parametric posterior probability that each cell is assigned a cell-type label. Thus, it allowed us to evaluate the label confidence for each cell and use label entropies to highlight transitions along cell developmental trajectories. Furthermore, Forest Fire Clustering enabled us to perform robust inductive inferences in an online-learning context and scale easily to millions of cells (35725981). To recover the dynamics from scRNA-seq data, we developed a neural-network-based ordinary differential equation called scDVF to elucidate the dynamics underlying scRNA-seq experiments by modeling the gene expression changes of single cells across time. Leveraging a deep learning model, scDVF can model non-linear, high dimensional gene interactions in single-cell dynamical systems. Furthermore, it allows us to perform in silico studies to explore dynamics of biological processes over time (36449617). To relate microbes to genes within heterogeneous RNA-seq data, we developed a method called LDA-link that identifies relationships between genes, microbes and cell types using reduced-dimensionality LDA topics. We used this tool to analyze bulk RNAseq data from 115 asthmatic patients with clinical information, microscope images and single-cell profiles and found 32 known microbe-gene connections among novel connections. The LAD-link represents a general strategy to deal with heterogeneous RNAseq data that is applicable to other sample types (32571363).

Overall Experience Working in a Consortium Framework. We have extensive experience in participating and leading large-scale analysis, developing QC metrics and uniform processing pipelines for consortia, such as ENCODE (22955616, 22955619), PsychENCODE (30545857, 26605881), the Extracellular RNA Communication Project (exRNA) (30956140, 30951667), GENCODE (22951037, 30357393), IGVF, HGSVC, PCAWG/TCGA (32025007), Kbase (29979655), 1000genomes, SCORCH and dGTEx. Here, we emphasize some examples to demonstrate our extensive experience in large-scale data analysis and building data infrastructure.

I) ENCODE, modENCODE, GENCODE and IGVF.

We worked extensively in the ENCODE and modENCODE GENCODE project. For instance, we 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, 22039215). We developed methods to construct and analyze human and model organism gene regulatory networks (22955619, 25164757, 22039215, 20439753, 21177976, 21430782) using ENCODE and modENCODE datasets. We also worked on an outgrowth of this project, GENCODE. Here we annotate pseudogenes (22951037) and participate in the annotation of the gnome (30357393). We also participated in the IGVF consortium as part of the Data Analysis Coordination Center (DACC). We are part of the efforts in providing the biomedical community with a catalog of human genetic variants and integrated information of their phenotypic impact within the consortium.

Figure 1. Overview of the ENCODE resource.

II) Brainspan and PsychENCODE

We participated in the Brainspan consortium and led the PsychENCODE data analysis. For the Brainspan consortium, we analyzed a variety of genomic data modalities including transcriptome, DNA methylation, and histone modification across multiple brain regions during development. We observed a widespread transcriptomic transition beginning during late fetal development (30545854). After that, Brainspan consortium has started work on the PsychENCODE project aiming to understand regulatory variants in the context of their functional connections to psychiatric disease. In the PsychENCODE consortium, we analyzed how proportions of cells contribute to variations in tissue-level gene expression across individuals (30545857). The base of the pyramidal resource is the datasets generated 

by PsychENCODE, including bulk transcriptome, chromatin, genotype, and Hi-C datasets and single-cell transcriptomic data from ~32,000 cells for major brain regions. Adding publicly available dataset, we first used Nonnegative Matrix Factorization to decompose bulk tissue data. Then we deconvolved the bulk tissue expression across 1,866 individuals in PsychENCODE and GTEx using single-cell data via non-negative least squares. We found that cell fraction changes contribute to >88% of the cross-population variation and the cell fraction changes were associated with aging and disorders.  At the top of the pyramidal resource, we also build a Deep Structure Phenotype Network (DSPN). Now we're continuing work on a new phase of the PsychENCODE project focusing on single cell analysis and on constructing a cell atlas of the prefrontal cortex.

III) ERCC

We also have experience in developing pipelines for analyzing exRNA-Seq data. As part of the exRNA consortium, we developed the extracellular RNA processing toolkit (exceRpt) pipeline for uniform processing throughout the consortium (30956140, 30951667), which has been used over 80,000 times. exceRpt performs sequential alignment of RNA to contaminants, to human transcriptome and genome, to human repetitive elements, and finally to exogenous sequences. exceRpt has uniformly processed and applied QC standards to all the datasets in the ERCC exRNA Atlas (http://exrna-atlas.org/) (26320941). The exceRpt pipeline enables a variety of user-specified customizations, including RNA reference prioritization, random barcoding, spike-in support, and detailed quantification reports. The pipeline is available as source (at github.gersteinlab.org/exceRpt) or wrapped in a user-friendly, browser-based interface available (at genboree.org).

Figure 2. Flowchart of exceRpt.

IV) Variant Consortia

We have extensive experience in identifying and interpreting genomic variants as part of the 1000 Genomes Project and the Human Genome Structural Variation Consortium (HGSVC) (https://www.internationalgenome.org/human-genome-structural-variation-consortium/). We led the functional subgroup of the 1000 genome consortium. For instance, We developed a variant-prioritization pipeline named FunSeq that included an adjustable data context (24092746, 25273974). This tool has been widely used to identify disease-causing mutations for further in-depth analyses to understand the mechanisms underlying disease pathogenesis. FunSeq links each noncoding mutation to target genes, and prioritizes such variants based on functional annotation, sequence features, conservation, network connectivity, and mutation frequency in diseases. We also developed Variant Annotation Tool (VAT) to annotate the impact of protein sequence mutations. VAT provides transcript-specific annotations of mutations according to synonymous, missense, nonsense or splice-site-disrupting changes (22743228). In addition to developing tools, we also analyzed 2,951 putative LoF variants obtained from 185 human genomes to determine their true prevalence and properties. We estimated that human genomes typically contain ~100 genuine LoF variants with ~20 genes completely inactivated. We identified rare and likely deleterious LoF alleles and described functional and evolutionary differences between LoF-tolerant and recessive disease genes, as well as a method for using these differences to prioritize candidate genes found in clinical sequencing studies (22344438). We have also been associated with Centers for Mendelian Genomics (CMG) (http://mendelian.org/), Genome-Sequencing-Program (GSP) (https://www.genome.gov/Funded-Programs-Projects/NHGRI-Genome-Sequencing-Program).



V) PCAWG/TCGA.

We worked on several cancer projects over the past years (32084333, 32025015, 32024998, 32024824). We are heavily involved in the Pan-Cancer Analysis Working Group (PCAWG)/TCGA (32025007), we were both co-leaders and participants in the working groups. We utilized comprehensive variant data from PCAWG to ascertain the molecular functional impact of each variant and showed that in addition to high- and low-impact mutations, there is a group of medium-impact putative passengers predicted to influence gene activity (32084333).

VI) Other Consortium Work.

We have also been associated with Kbase (29979655), SCORCH Consortium (https://www.igs.umaryland.edu/topics/scorch/), dGTEx Consortium (https://dgtex.org/), Northeast Structural Genomics Consortium (https://www.nesg.org/), North East Biodefense Center (<http://www.nbc.columbia.edu/>).

Machine learning models. We have much experience in data integration and machine learning. For example, we applied topic modelling to identify gene groups that show significant co-expression behaviors and related functions (32657410). We further extended the approach to integrate multiple data sources including gene expression and microbiome (33059594). Moreover, we developed many tools for comparative gene regulatory network analyses, including Loregic (25884877), which analyzes regulatory cooperativity; OrthoClust (25249401), which discovers novel human gene functions via clustering cross-species gene co-expression networks; and DREISS (27760135), which analyzes the dynamics of gene regulatory networks using dynamic models.

Recently, we have done some studies that apply deep learning methods to biomedical data. Especially, we have experience in designing deep learning methods according to the nature of the dataset and the problem. The interpretability of the models is considered important when we design the models, because that enables novel biological insights from the trained models. We proposed several novel neural network models in recent publications. For instance, (1) a convolutional neural network (CNN) model for the refinement of regulatory region boundaries, for which deployed model interpretation methods were also developed to identify features that are important for the prediction (34252960). (2) In another work, we developed a 3D CNN to predict thermodynamic stability changes caused by mutations from 3D structural data (33253214). (3) Another example is a recurrent neural network (RNN) based model for the prediction of the alternative splicing events from RNA sequences and RBP binding (32584815). (4) We developed an interpretable Deep Structured Phenotype Network (DSPN) (30545857), a deep learning model based conditional Deep Boltzmann Machine architecture with multiple layers, which improved GWAS disease prediction by sixfold compared to additive polygenic risk scores. DSPN highlights key genes and allows imputation of missing transcriptome from genotype.

Physical modeling. We have extensively employed physical modeling to understand conformational changes, physical properties, structure and functions of protein upon genetic mutations in the past. For example, we examined localized perturbations of proteins that may severely affect protein stability and function by calculating changes in localized frustration indices, which is indicated by increase or decrease in unfavorable interactions with neighboring residues after introducing SNVs (27915290). In another work, we made improvements on the database of molecular motions by relating protein motion classification to function classification through GO terms, by linking motions to active sites, and by generating more realistic interpolations between two protein structures to study changes in protein conformations (12520056)*.* In another example, we investigated the conservation of potential allosteric residues across and within species through identifying allosteric hotspots and provided a platform for the study of general trends and conservation patterns in human genomics (27066750). Moreover, we presented a dynamics-based search method of mutational hotspot communities, which significantly increased the sensitivity for cancer driver detection *(*31462496). Besides researching protein dynamics and conformational changes, we also studied physical properties and structural features of protein. By calculating packing efficiency during conformational change, we offered insights into protein packing in the context of secondary structure, protein cores and surfaces (19472340).In addition, we studied the packing at the protein-water interface to shed light on surface protein behaviors in the context of solvent (8816770).

Moreover, we have had experiences combining physical and statistical modeling. For example, we developed a supervised machine learning method, GenoDock, to investigate how a given variant impacts the binding affinity of protein-drug co-crystal structures (31279629). By mapping SNVs to protein structures with drug ligands, we linked genomics and 3D protein structures through the calculation of binding affinity, shedding light on the disruptive role of a point mutation on drug resistance (31279629).

LoF annotation and pseudo genes. Our lab has substantial experience in annotating variants that result in the Loss-of-Function (LoF) of human genes by integrating variant calls (vat.gersteinlab.org) and genome annotation with data from online resources such as the protein family database PFAM (33125078), functional and structural domains databases SMART (29040681) and SCOP (24293656). We previously developed a pipeline called Annotation of Loss-of-Function Transcripts (ALoFT) (28851873), that provides comprehensive annotation of putative loss-of-function (pLoF) variants (22344438). Using the annotation output of ALoFT as predictive features, we further inferred the pathogenicity of pLoF variants in Mendelian diseases, autism, and cancer. ALoFT allows for the identification and prioritization of high-impact putative disease-causing pLoF variants in individual genomes.



**Figure 5. ALoFT classification of 1000 Genomes and HGMD variants.**

Pseudogenes are disabled counterparts of functional protein coding genes. Over the years, using a combination of manual curation and computational analysis, we have identified and characterized pseudogenes in human (22951037,30357393) and multiple species (25157146,19123937). A close inspection of pseudogene disabling mutations reveals that the pseudogenization process is closely related to LoF events such as premature truncation of proteins, disruption of splicing and structural domains (24026178,22344438,21205862). And the pseudogenization of a gene undergoes fixation in the population if the mutations at the locus which are freed from selection pressure. Thus, the pseudogenes could be considered as fixed LoF variants (20210993). There are over 14,000 pseudogenes in the human genome (22611337), which can be used as important resources to characterize the LoF alleles.



**Figure 6. pLoFs in last exons.**

Network analysis. We have extensive experience in biological network science. We have developed a network analysis platform (http://networks.gersteinlab.org) which encompasses tools to determine small-scale network motifs, such as feed-forward loops and feedback loops, and large-scale structures, such as overall network hierarchies and bottlenecks of networks. These have been published in numerous analyses including identifying enriched network motifs with Loregic (25884877), cross-species network clustering with OrthoClust (25249401) and calculating the impact of conserved or species-specific regulatory networks on gene expression with DREISS (27760135). Additional network analysis methods have been formulated in a web-accessible network toolkit called TYNA (17021160). This platform has been used to analyze the human regulatory network, the network associated with cancer, the phosphorylation network in yeast, the yeast regulatory network, and other model organism networks (25880651). The Gerstein lab has previous experience in mining the yeast regulatory network (18451266, 17690298), finding the occurrence of various motifs used by the phosphorylation network and the transcriptional regulatory network. We also published a review on how network analysis is a great unifier in biomedical data science (https://doi.org/10.1146/annurev-biodatasci-080917-013444). Moreover, constructed networks can help infer the direct and indirect effects of genomic variants. We have made use of network properties such as centrality to evaluate the functional significance of genomic variants, revealing that the correlation between them depends on the network type (23505346). Network hierarchy represents another useful concept for predicting the impacts of genomic variants, with the perturbation of elements at the upper layers (such as master regulators) causing more widespread effects than the perturbation of elements at the lower layers, which are more localized. We have developed different methods for determining network hierarchies, such as the HirNet method, which was applied to compare the hierarchies of the phosphorylation and TF binding networks (25880651). Genomic variants can also lead to disruptions of network connections. As a result, some recurrent patterns may be perturbed, such as TFs that frequently co-regulate target genes. We have developed the DiNeR method for identifying such changes and analyzing their consequences to downstream gene expression programs (32615918). On a larger scale, some network perturbations may propagate to cause major network rewiring. We have developed the TopicNet method to measure such rewiring in transcriptional regulatory networks (32657410). We have also applied this idea to study network rewiring in cancer cells, as part of our efforts in producing a general resource for cancer research based on ENCODE data (32728046).