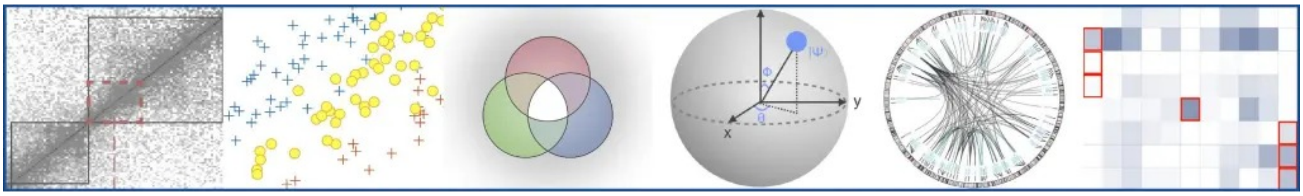
GERSTEINLAB

Research

Soon, sequencing one's genome may become as commonplace as getting an X-ray. Consequently, personal genomes will increasingly serve as the lens through which the public views biology. Appropriately interpreting personal genomes – particularly in relation to disorders such as cancer and neurological diseases – is therefore of key importance. Moreover, the expansion of DNA sequencing and other data generating technologies related to images, structures and sensors is making biomedical data science a growing area with broad connections to other data-intensive disciplines with the umbrella of data science. Significant advances in computing have paralleled the ongoing rapid growth in biomedical data generation, giving rise to new approaches in machine learning, network science, and physical modeling. The Gerstein lab acts as a connector, bringing quantitative approaches from disciplines such as computer science and statistics to address practical questions and large-scale data in molecular biology. Often, we carry out our work in multi-disciplinary teams through collaborative efforts (e.g. in consortia such as PsychENCODE, ENCODE, GSP/CMG and 1000 Genomes). Below, we describe specific aspects of our research that work toward our overall goal of interpreting personal genomes and advancing biomedical data science as a field.



Genome Annotation

At the heart of our lab is human genome annotation. We have made significant efforts to annotate the human genome, through active participation in worldwide collaborations including ENCODE, modENCODE, and GENCODE, as well as through the development of computational tools. Our work targets coding and noncoding genomic regions and ranks somatic and germline variants in relation to their potential functional impact and deleteriousness in causing disease.

Our main focus has been on transcription factor binding sites and non-coding RNAs (ncRNAs). We have developed tools that leverage comprehensive CHIP sequencing (CHIP-seq) data to detect transcription factor binding sites and utilize this type of information to predict the expression of target genes. Other tools that we built identify ncRNAs and regions of intragenic transcription by processing datasets from RNA-seq and CHIP-seq assays. Additionally, we have contributed to the annotation of pseudogenes. See encode/annotation and pseudogene papers.

Disease Genomics (Neurogenomics & Cancer Genomics)

The declining cost of next-generation sequencing has allowed researchers to rapidly study the genomic contributions to disease. In the Gerstein Lab, we have contributed to this effort through comprehensive studies and computational tools that aim to establish connections between genomic variants and disease. We have studied a considerable number of diseases with a focus on cancers and brain disorders. Recent efforts include developing tools to prioritize noncoding driver mutations in cancer, integrating genome annotations with cancer genomes to develop a resource for cancer genomics, and studying the effects of nondriver mutations in cancer. In tandem, we co-led an effort to establish a comprehensive functional genomic resource that pertains to the human brain, which involved integrating data at the single-cell level with plentiful bulk functional genomics data. See neurogenomics and cancer genomics papers.

Personal Genomics & Genomic Privacy

For specific personal genomes, we have developed various “callers” to find variants. Comparing variant calls between individuals shows that all humans share the vast majority of their genomes, yet a small fraction of each individual’s genome sequence shapes her or his unique combination of physical and physiological traits. We have developed tools that study personal genomics and link molecular phenotypes such as gene expression to differences in parental alleles.

Overall, this work reveals the potential for high-dimensional genomic data to reveal sensitive personal information such as disease states. Using information theory and other approaches, we have developed tools to assess the feasibility of sharing molecular data without jeopardizing the privacy of sample donors. See personal genomics and privacy papers.

Data Science and Biological Networks

We have developed tools to build and analyze multi-omics, regulatory networks, protein-protein interactions and metabolic pathways, identifying key nodes such as hubs and bottlenecks. We have also integrated networks with dynamic gene-expression data (identifying transient hubs), three-dimensional protein structures, and other regulatory data to find large-scale regulatory principles for biological systems. Finally, people have better intuition for commonplace networks – such as those in social and computer systems – compared to biological networks. Thus, we have found that cross-disciplinary comparisons are helpful to elucidate system-level properties of biological networks, such as the association of greater connectivity with more evolutionary constraints. See data science, network, and bioinformatics tools papers.

Macromolecular Motion & Dynamics

While non-coding regions play an important, if underappreciated, role in genome function and disease, we also characterize coding sequences and drill deep into their protein products. By analyzing protein motions, we can better predict how a mutation affects function. This effort involves devising a system for characterizing motions in a standardized fashion in terms of key statistics, such as the degree of rotation around hinges. Our approach is guided by the fact that protein mobility is highly restricted by tight packing. We have developed a variety of tools to analyze protein structures and motions, including measuring packing efficiency using specialized geometric constructions (e.g., Voronoi polyhedra). Recently, we applied a combination of molecular motion simulations and network analyses to identify cancer mutation hotspots within proteins. See: molecular motion and structure papers.

Interpretable Machine Learning Tools

The rapid increase in biomedical data during the last two decades has also engendered the need for artificial intelligence tools that can find patterns embedded in large-scale datasets and study a variety of data representations. In machine learning – a branch of artificial intelligence that integrates algorithmic and statistical techniques – we have developed tools to perform predictive tasks that provide insights on genomic research. We focus on approaches that are “interpretable” in that they have a clear biological or physical basis. Our tools process large-scale datasets to, for instance, functionally prioritize genomic variants with respect to their biological function or potential contribution to disease or predict protein binding. See Gerstein Lab repository on GitHub for more details.

References

See Papers.GersteinLab.org — in particular, Best Papers and listing of Key Contributions.

Some talks giving a quick overview of the lab: 5′ animation ('20), 15′ powerpoint ('19)

More information on research interests can also be found here.