Gerstein Lab Experience in Computational Tools and Databases

We have taken on crucial roles in several high-profile NIH-funded projects, including ENCODE (Encyclopedia of DNA Elements) [1][3], psychENCODE [2-5], the 1000 Genomes Project [4-7], and dGTEx. We have been at the forefront of developing essential computational tools and databases that have advanced the field of genomics. Notable contributions include: Database of Macromolecular Motions, a comprehensive resource categorizing macromolecule conformational changes [6][8]; tYNA, a tool for analyzing molecular networks that aids in understanding complex interactions within biological systems [8][9]; PubNet, a publication network analysis tool that helps visualize the relationships and influences within scientific literature [8][10]; PeakSeq, a tool for identifying regions in the genome bound by specific transcription factors, crucial for understanding gene regulation [9][11]; CNVnator, a tool that categorizes block variants in the genome, facilitates studying structural variations and their implications in disease [10][12]. We also develop sophisticated data science methodologies for interpreting molecular networks [11][13], evaluating genomic privacy [12][14], and conducting integrative data mining [13][15].

Detailed experience encompassing the following key areas is provided below:

1. Privacy and security
2. Brain system biology with data resource and regulatory network
3. Asthma and respiratory disease
4. Significant genomic databases
5. General Tool Development
6. Biological networks and analysis

References

1. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91-100.
2. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362(6420).
3. Emani PS, Liu JJ, Clarke D, Jensen M, Warrell J, Gupta C, et al. Single-cell genomics and regulatory networks for 388 human brains. bioRxiv. 2024.
4. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
5. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. 2011;39(16):7058-76.
6. Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, et al. The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res. 2006;34(Database issue):D296-301.
7. Yip KY, Yu H, Kim PM, Schultz M, Gerstein M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics. 2006;22(23):2968-70.
8. Douglas SM, Montelione GT, Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome Biol. 2005;6(9):R80.
9. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66-75.
10. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974-84.
11. Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, et al. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol. 2011;7(11):e1002190.
12. Gürsoy G, Emani P, Brannon CM, Jolanki OA, Harmanci A, Strattan JS, et al. Data Sanitization to Reduce Private Information Leakage from Functional Genomics. Cell. 2020;183(4):905-17.e16.
13. Bertone P, Kluger Y, Lan N, Zheng D, Christendat D, Yee A, et al. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res. 2001;29(13):2884-98.

1. **Privacy and security**

With the recent advancement of machine learning and deep learning methods, the analysis of genomic data has reached a new era. A large-scale functional genomic dataset will increase the statistical power of models trained on it. As more participants are required in such kind of study, the privacy of these participants should be maintained in the research process. Unlike traditional genomic data where the research community has already recognized the risk of leaking sensitive information, functional genomics data, such as RNA-seq or ChIP-seq, could pose serious privacy risks that people previously were unaware of [PMID 26828419][PMID 21134889]. Prof. Gerstein's genomic privacy work, especially assessing risk factors related to genomic privacy ensures the protection of personal data. We have observed that in a phenotype-genotype dataset, one can connect seemingly independent phenotype and genotype entries and leak sensitive personal genomic information with various linking attack schemes. We could infer someone's eQTL status from gene expression and yielded more than 90% accuracy when applied to the gEUVADIS dataset. Also, through statistical analysis, we provided a quantitative measurement of information leakage and predictability of the genotype [PMID 26828419]. In another work, we assessed privacy leakage from functional genomic data, more specifically RNA-seq data, and revealed that genomic deletion could be detected from raw sequencing data of large study cohorts such as GTEx and ENCODE. Thus, the raw data of these cohorts could also be susceptible to linking attacks and leaking important private information. We also proposed systematically removing the dips in the RNA-seq data profile as an anonymization method to protect genomic privacy in this type of functional genomics dataset [PMID 29934598]. We also developed FANCY, a software used to estimate the number of leaked variants from RNA-seq, ATAC-seq, and ChIP-seq datasets. We reached an average R2 of 0.95 on each independent test set and were able to use FANCY to accurately predict low numbers of leaking variants. An accurate estimate of leaked variants from these functional genomics datasets could allow scientists to evaluate privacy risks associated with certain datasets [PMID 32726397].

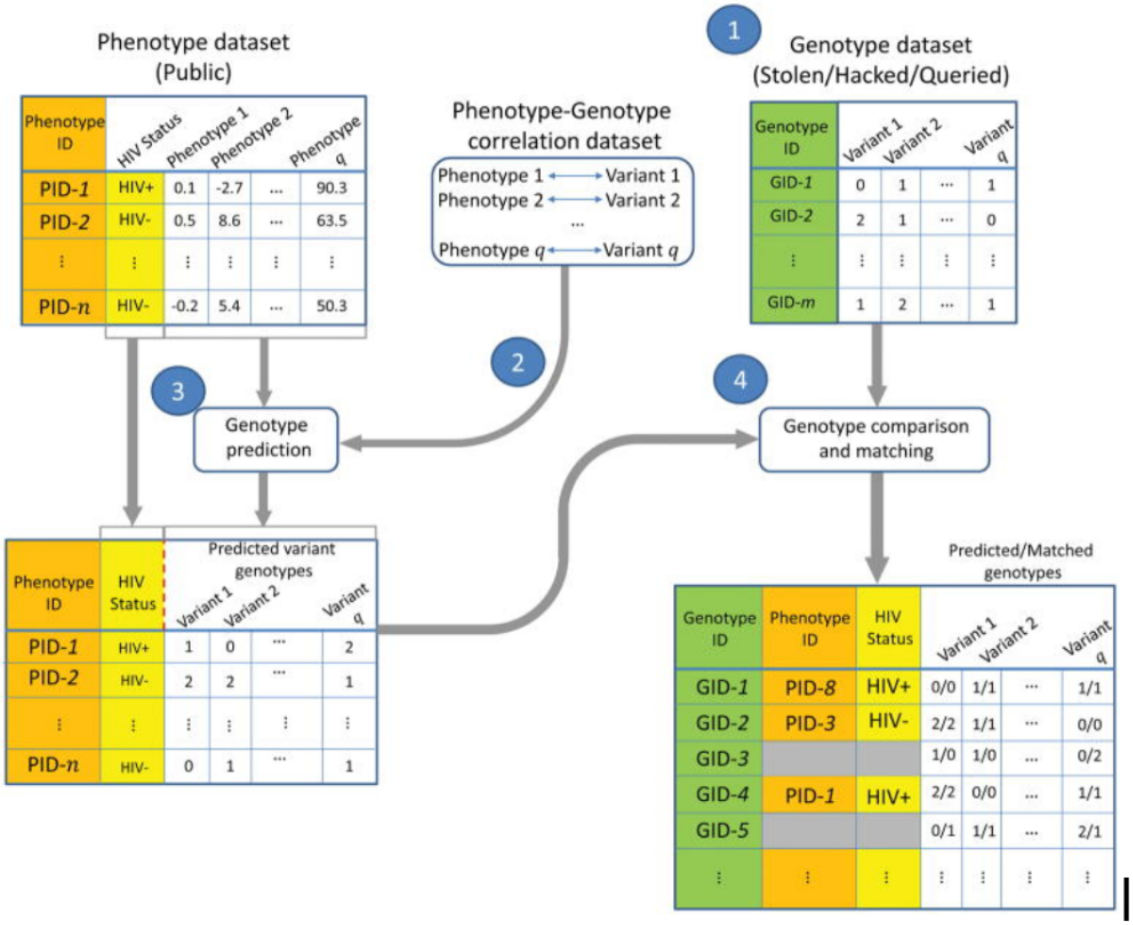

Fig 1. A graph summary of linking attack schemes with gene expression and genomic variants

On the other hand, to alleviate privacy concerns on genomic datasets and balance between protecting genomic privacy and the utility of the dataset, Prof. Gerstein's commitment to genomic privacy pertains to data protection. We have developed a series of algorithms and file formats to promote the proper sharing and analysis of data while protecting genomic privacy. RSEQtools is a suite of tools to analyze functional genomic data in the format of MRF (Mapped Read Format), which was proposed by us to enable the anonymization of confidential sequence information while allowing researchers to still be able to conduct many useful studies. MRF is a compact format where only minimal useful information is provided for analysis without revealing sensitive private information on the sample. With the help of RSEQtools, one can construct an efficient analysis pipeline on functional genomic datasets to quantify gene expression, visualize expression profiles, and identify transcriptionally active regions [PMID 21134889]. We also proposed a data sanitation method and implemented it as pTools. It could convert normal BAM files into a privacy-protecting pBAM format and save the difference between pBAM and real BAM files in a different .diff file. pBAM files were readily available for further analysis of the functional genomic data, while the real BAM file and .diff file could be kept as controlled access from the consortium side [PMID 33186529].

Homomorphic encryption is a powerful tool to help ensure the safety and privacy of datasets in machine learning models. Various machine learning tasks could be performed on encrypted data with relatively high accuracy, and applying homomorphic encryption on genomic machine learning

models could help prevent leakage of key information. Prof. Gerstein's work on homomorphic encryption machine learning models helps to increase the security of machine learning applied to sensitive genomic information. Various machine learning methods could achieve genotype imputation, but the need for genotype information could lead to serious leakage and privacy risks. We developed a privacy-preserving statistical inference algorithm called p-Impute, which was based on homomorphic encryption, namely the BFV encryption scheme from SEAL. We overcame difficulties related to optimization and performance and were able to impute unknown genotypes with an accuracy comparable to state-of-the-art non-encrypted methods [PMID 34758288]. We also explored homomorphic encryption machine learning models on cancer type prediction and also achieved great performance on large-scale studies like TCGA. The somatic mutations in the original dataset, including CNVs and SNVs, were encrypted and the model performed well even compared with state-of-the-art models while also preserving sensitive patient genomic information [PMID 36717667].

Recently, blockchain has received widespread attention. Due to its properties of security, immutability, transparency, and decentralization, it is regarded as one of the most promising methods to protect the safety of the data and access logs as transactions. This method would be ideal for storing sensitive biological information, and Prof. Gerstein's experience in this field helps to protect sensitive genomic information. We developed an Ethereum smart contract for storing and querying pharmacogenomics data. A common challenge of applying blockchain to everyday problems was its inefficiency in storing and querying data. However, we addressed this problem with special designs on data structure and algorithm to enable time and memory-efficient data insertion and querying [PMID 32487214]. In another work, we demonstrated an efficient method of storing and querying genome dataset access logs called MultiChain. We optimized the method for efficient data storage and query, and it was selected as the winner of the iDASH competition at a workshop held in San Diego, CA in October 2018 [PMID 32693796]. We recently created a blockchain framework where sensitive genomic information, like VCF files and SAM files, could be stored and accessed on a blockchain. The data itself could be queried at a higher efficiency than normal blockchain methods, and we provided an analysis toolkit for VCF files (VCFtool) and SAM files (SCtools) on the blockchain so that the analysis could be accomplished securely and efficiently. We envisioned that anyone could create their own blockchain storing their genomic information and safely share it with their health providers or any other entity [PMID 35765079].

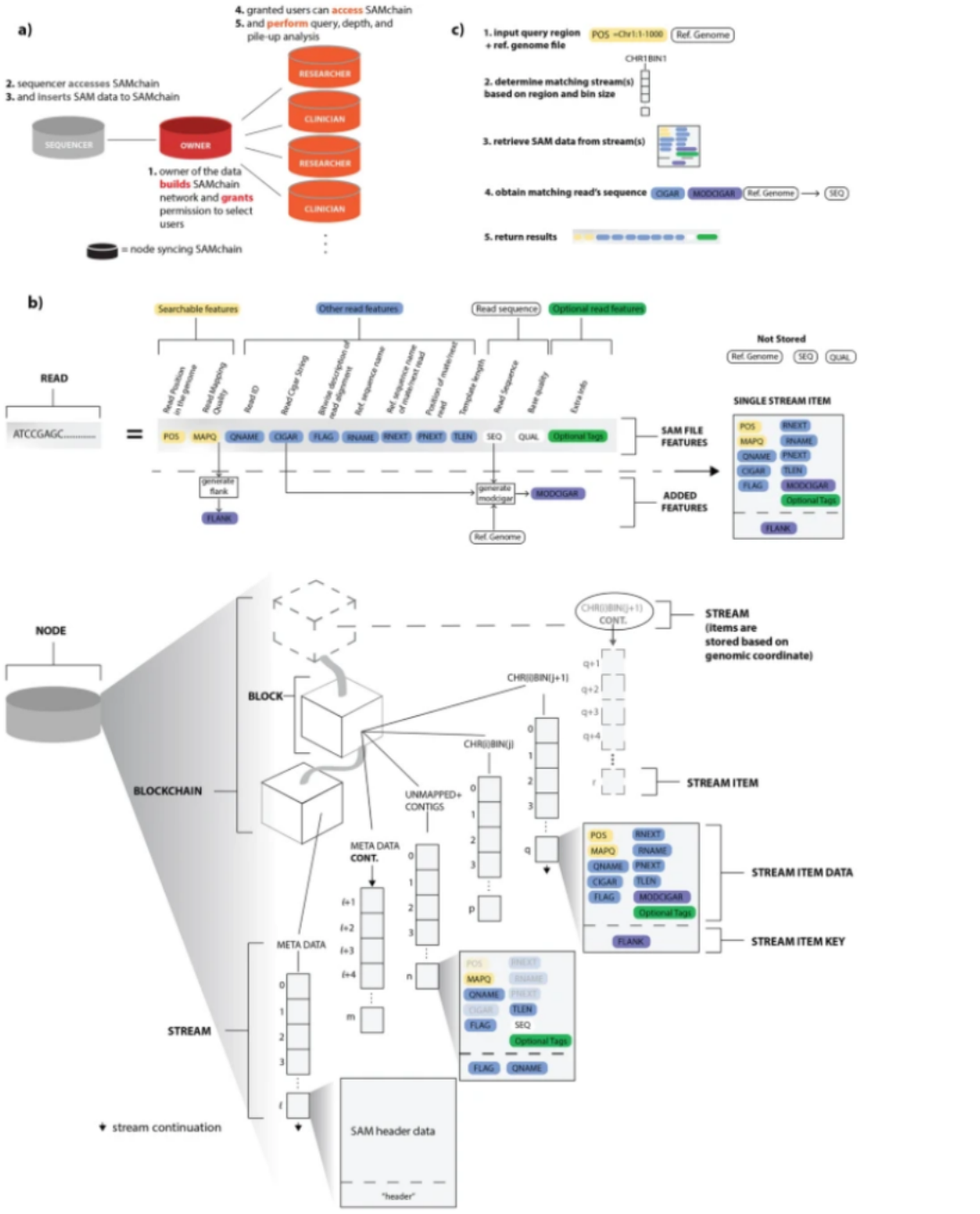

Fig 2. Framework and design of SAMchain and VCFchain

2. **<u>Brain system biology with data resource and regulatory network</u>**

Gerstein lab, as part of the PsychENCODE Consortium, played an essential role in the data analysis. The PsychENCODE Consortium has generated a comprehensive online resource and integrative models for the functional genomics of the human brain [PMID: 30545857]. The resource forms a three-layer pyramid. 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. We have merged these with data from Genotype-Tissue Expression (GTEx), ENCODE, Roadmap Epigenomics, and single-cell analyses. Via uniform processing, we created a harmonized resource, allowing us to survey functional genomics data on the brain over a sample size of 1866 individuals. These data represent a range of psychiatric disorders, including SCZ, bipolar disorder, and autism spectrum disorder.

From this uniformly processed dataset, we created derived data products. These include lists of brain-expressed genes, coexpression modules, and single-cell expression profiles for many brain cell types; ~79,000 brain-active enhancers with associated Hi-C loops and topologically associating domains; and ~2.5 million expression quantitative-trait loci (QTLs) comprising ~238,000 linkage-disequilibrium–independent single-nucleotide polymorphisms and of other types of QTLs associated with splice isoforms, cell fractions, and chromatin activity. The derived data also enables comparison between the brain and other tissues.

The top level of the resource consists of integrative networks for regulation and machine-learning models for disease prediction. The networks include a full gene regulatory network (GRN) for the brain, linking transcription factors, enhancers, and target genes from merging of the QTLs, generalized element-activity correlations, and Hi-C data. We then embedded the regulatory network into a deep-learning model, the DSPN, to predict psychiatric phenotypes from genotype and expression. The model has improved prediction accuracy over traditional additive models. In addition to trait prediction, it can highlight key genes and pathways associated with disorder prediction, including immunological, synaptic, and metabolic pathways, recapitulating de novo results from more targeted analyses.

For the aforementioned data, we can integrate them into the system biology data platform. This will encompass a comprehensive array of elements such as brain-active enhancers, sets of Hi-C linkages, and topologically associating domains; single-cell expression profiles for many cell types; eQTLs and further QTLs associated with chromatin, splicing, and cell-type proportions; and GRN for the brain. Furthermore, we will develop a cloud-based psychiatric phenotype prediction tool while prioritizing associated genes and networks. These resources can aid researchers in deciphering the molecular mechanisms within the brain, and proposing potential avenues for future research into the etiology of brain disorders.

One key aspect of the human brain comprehensive functional genomics resource generated by the PsychENCODE Consortium was mapping interactions and regulatory relationships across different omics data types.
We used standardized pipelines like the ENCODE/GTEx pipelines for uniform processing of RNA-seq, ChIP-seq, and other genomic data across datasets. This allowed consistent comparison across tissues and integration of data from different sources. Specifically, to identify brain transcriptional signatures, we applied standardized and established ENCODE pipeline to uniformly process RNA-seq data across PsychENCODE, GTEx, and Roadmap datasets. This consistency ensured our expression data and downstream analyses comparable to previous work. Using these data, we identified noncoding regions of transcription and sets of differentially expressed and coexpressed genes [PMID: 30545856]. For eQTL mapping, we closely followed

the standard GTEx pipeline [PMID: 25954001] and identified substantially more eQTLs (2.5 million) associated with around 33,000 expressed genes in PFC than previous studies, leveraging our large sample size. For enhancer identification, we used standard ENCODE ChIP-seq pipelines to process data, ensuring maximal compatibility of the results. As a result, we annotated a reference set of ~79,000 brain enhancers in the prefrontal cortex (PFC).

We further integrated the genomic elements into a regulatory network. We first processed adult brain Hi-C data, identifying 2735 topologically associating domains (TADs) and approximately 90,000 enhancer-promoter interactions. As expected, the majority of interactions occur within the same TAD, and the genes with more enhancers tend to have higher expression levels. Upon incorporating QTLs, we observed that QTLs linked distally to expressed genes (eGenes) via Hi-C interactions had significantly stronger associations compared to those with SNPs directly within eGene promoters or exons. Additionally, we explored the regulatory connections between transcription factor (TF) activity and target gene expression using elastic net regression. Overall, our work generated a comprehensive regulatory network linking enhancers, TFs, and target genes. This network includes 43,181 proximal and 42,681 distal connections involving 11,573 protein-encoding target genes. The proximal connections TFs to target genes via promoters, whereas the distal connections link TFs to target genes via enhancers [PMID: 25053837]. We also generated potential cell-type-specific regulatory networks. In these networks, several well-known TFs were identified for association with brain development, such as NEUROG1, DLGAP2, and MEF2A for excitatory neurons and GAD1, GAD2, and LHX6 for inhibitory neurons [PMID: 10640277; PMID: 28870203; PMID: 2069816; PMID: 17376969]. These networks not only enhance our understanding of brain development but also serve as a valuable resource for further research. This resource was published in the journal Science in 2018.

We can implement these pipelines to process the data sets from various AMP programs, as well as publishing them in established journals and making them accessible to researchers via a cloud-based platform. We will also implement any other standard processing pipelines that are needed for the AMP programs for other genomics data.

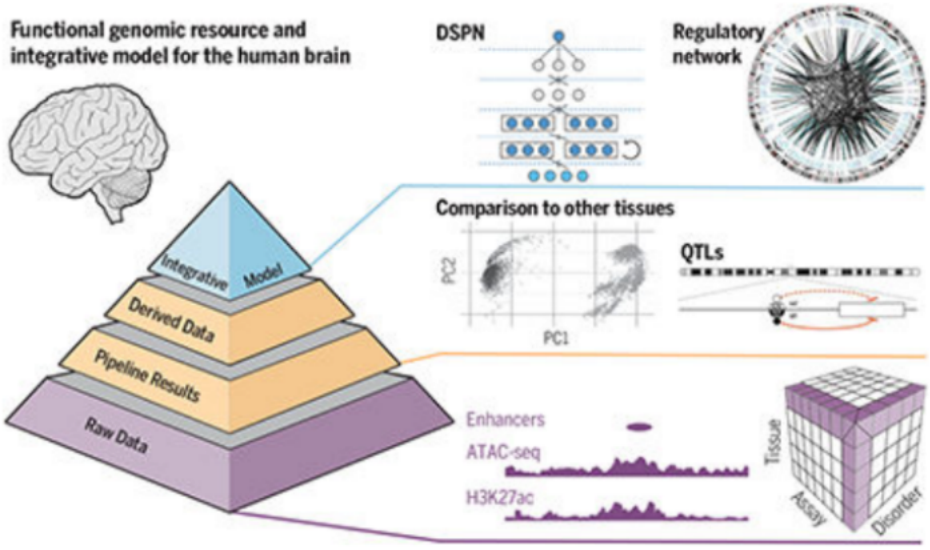

Fig 3. A comprehensive functional genomic resource for the adult human brain

Single-cell genomics has revolutionized the understanding of complex biological tissues like the brain. Despite advancements, a significant gap remained in linking genetic variants to their effects on gene expression at the cellular level. Traditional genomics studies often used bulk tissue data, which could not specify cellular contexts or reveal the cell-specific impact of genetic variants. Addressing this need, the work from Prof. Mark Gerstein's lab sought to elucidate how genetic diversity influences gene expression across different cell types in the human brain, particularly focusing on neuropsychiatric disorders [PMID: 38562822].

We processed single-nuclei multi-omics datasets into a significant resource, containing data from over 2.8 million nuclei derived from the prefrontal cortex. This resource spans 28 distinct cell types, allowing detailed analysis of gene expression and chromatin states across a diverse set of individuals, including those with neuropsychiatric disorders. By analyzing the data, we identified over 550,000 cell-type-specific regulatory elements and more than 1.4 million single-cell expression-quantitative trait loci (eQTLs). These findings provide insights into the regulatory architecture of the brain at a resolution that was previously unattainable.

We constructed detailed GRNs for each cell type by integrating information from the identified eQTLs and the cell-type-specific regulatory elements. These networks elucidate the interactions between transcription factors, enhancers, and target genes, offering a dynamic view of gene regulation across different cell types and conditions. Leveraging the regulatory and expression data, we developed networks depicting cell-to-cell communication within the brain. These networks reveal how cells interact molecularly, shifting the understanding of cellular interactions in both healthy and diseased brain states. Besides, an integrative model Linear Network of Cell Type Phenotypes (LNCTP) was developed to simulate and predict the effects of genetic perturbations on cell-type-specific gene expression. This model aids in prioritizing disease-risk genes and potential drug targets, significantly enhancing the translational impact of their findings. All sequencing data, derived analysis files, and computational codes have been made available through the brainSCOPE resource portal (http://brainscope.psychencode.org/). This portal serves as a valuable tool for the broader scientific community, facilitating further research and validation of the findings.

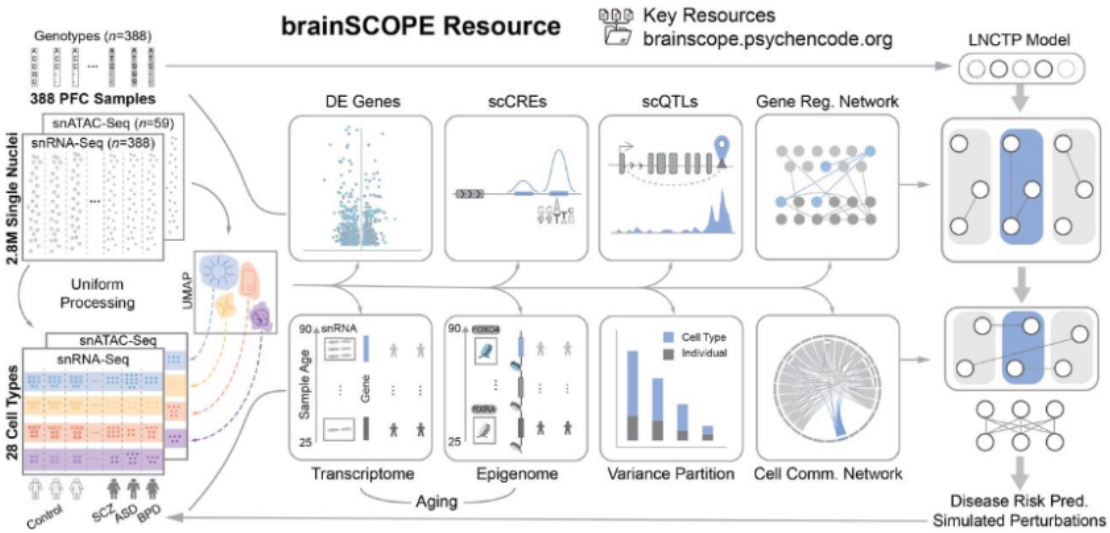

Fig 4. Brain Single-Cell Omics for PsychENCODE Resource

3. **Asthma and respiratory disease**

The Gerstein laboratory has developed innovative tools and methodologies to probe the microbial landscape in respiratory diseases and viral infections. Specifically, the Gerstein laboratory has created tools which connect microorganisms to human genes and tissues in the context of asthma and respiratory diseases, like SARS-CoV-2. These tools include platforms for extracellular RNA profiling, pipelines, and statistical modeling approaches, all of which have allowed researchers novel insight into host-microbe associations.

The analytic platform, exceRpt, offers a standardized framework for profiling extracellular RNAs to facilitate the identification of biomarkers and molecular signatures [PMID: 30956140]. exceRpt preprocesses, aligns, and quantitates small and long RNA-seq datasets and outputs both the respective sequences found in the sample and discovered exogenous microbes. With the identified microbes, exceRpt also has the capability to generate phylogenetic trees to support further analysis [PMID: 30956140]. The Gerstein laboratory has experience not only in profiling microbes, but also in connecting microbes to genes using a pipeline called LDA-link [PMID: 32571363]. LDA-link is a pipeline which integrates dimensionality reduction and statistical modeling techniques. Utilizing Latent Dirichlet allocation (LDA)-link methodology, the pipeline connects microbes to genes using reduced-dimensionality LDA topics. LDA-link has been applied to sputum samples from asthmatic patients, revealing both known and novel relationships between microbes and genes, thereby contributing to a deeper understanding of the microbial component in asthma pathogenesis [PMID: 32571363]. The Gerstein laboratory has expanded on the types of data they can integrate together with MLCrosstalk (multi-layer crosstalk). MLCrosstalk is a statistical modeling approach which integrates data from human genes, microRNAs, protein-protein interactions, and microbes to uncover the complex interactions that may occur in disease (PMID: 37410793). In the context of SARS-CoV-2, MLCrosstalk revealed a correlation between the abundance of specific microbes (*Rothia mucilaginosa* and *Prevotella melaninogenica*) and abundance of the virus, as well as associations with specific genes and pathways [PMID: 37410793].

Together, exceRpt, LDA-link, and MLCrosstalk form a robust toolkit for dissecting microbial contribution in diseases and disorders. Leveraging the versatility and efficacy of these tools, researchers can apply them to other datasets, including those from the NIH's Accelerating Medicines Partnership (AMP), to deepen our understanding of microbial dynamics across diverse diseases and disorders. By harnessing these innovative tools in conjunction with large-scale datasets, we can unlock new avenues for research, paving the way for targeted therapeutic interventions and precision medicine strategies tailored to disease.

4. **Significant genomic databases**

Dr. Gerstein's team has developed a comprehensive suite of genomic tools and databases that significantly contribute to major human genome projects, including the Encyclopedia of DNA Elements (ENCODE) [PMID: 32728249], model organism ENCODE (modENCODE) [PMID: 21177976], the GENCODE gene annotation project [PMID: 30357393], the EN-TEx [PMID: 37001506], the Developmental Genotype-Tissue Expression(dGTEX), and Impact of Genomic Variation on Function (IGVF) Consortium. These tools and data sources enhance our understanding of the genomic architecture by analyzing both protein-coding and non-coding regions. The team's long-standing expertise in developing methodologies to prioritize variants at multiple levels has led to significant advancements in genomic research.

Protein-coding variants. We have developed a variety of tools that prioritize protein-coding variants. Among them, **ALoFT** provides extensive annotations to putative loss-of-function variants (LoF) in protein-coding genes [PMID: 28851873]. Loss-of-function variants have attracted interest in clinical genetics because they can have a profound impact on gene function, often leading to disease in individuals who carry them, yet they can also be surprisingly prevalent in healthy individuals. ALoFT, which stands for "Annotation of Loss-Of-Function Transcripts," offers insights into the functional, evolutionary, and network features of these variants. By using ALoFT, we have been able to differentiate between LoF mutations that are deleterious in heterozygous states from those that may cause disease in the homozygous state. We have applied this tool to analyze the pathogenic potential of LoF variants, aiding in the study of their role in Mendelian diseases, autism, and cancer [PMID: 22344438]. The use of ALoFT in such analyses underscores its value in interpreting the clinical significance of genetic variations in a range of health conditions. We have also built a workflow based on **Frustration** to identify mutations that affect allosteric hotspots in proteins and identify key functional protein regions prone to genetic alterations [PMID: 27915290]. Notably, many disease-associated variants at predicted allosteric sites have historically been poorly understood, and the rarity of these variants complicates their analysis using conventional phenotype-genotype associations. To address this challenge, our workflow employs localized frustration as a metric to quantify and analyze the unfavorable local interactions caused by SNVs, providing crucial insights into their specific functional effects on protein structures (**Fig. 5**). Using this workflow, we analyzed data from the Protein Data Bank (PDB) and uncovered that disease-associated variants lead to greater localized frustration than non-disease variants, with rare variants causing more severe disruptions in local interactions.

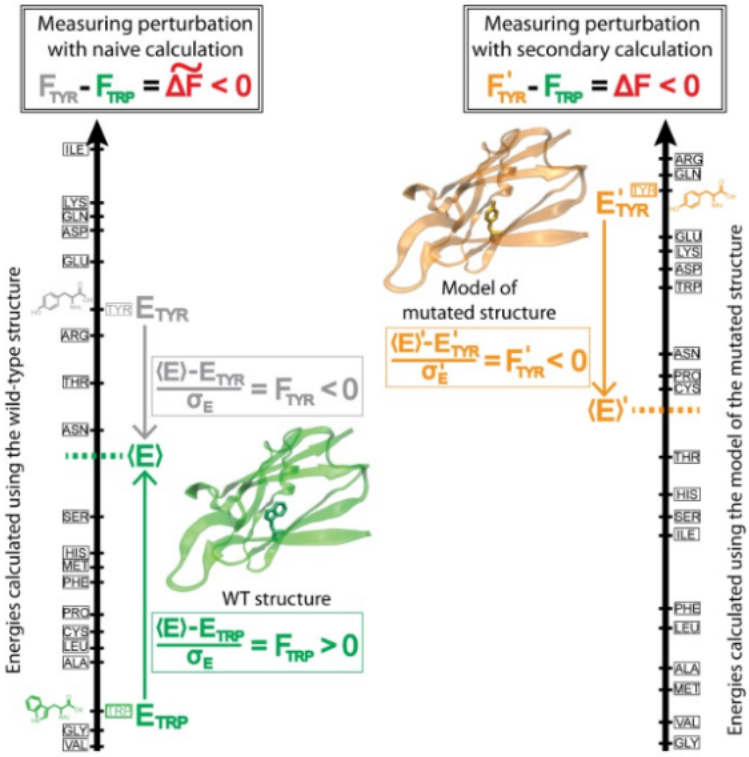

Fig 5. Prioritizing the effect of SNVs based on changes in localized perturbations (as measured by frustration).

Non-coding variants. In addition to coding variant prioritization, we have extensively analyzed patterns of variation in non-coding regions and their coding targets [PMID: 21596777, 22955619, 22950945], integrating these variants with biological networks and other features for prioritization [PMID: 23505346]. **FunSeq** is a comprehensive prioritization pipeline that integrates multiple methods to analyze variation in non-coding regions and their coding targets [PMID: 24092746], linking non-coding mutations to their relevant genes and prioritizing variants based on network connectivity and the potential disruptiveness of these variants. The tool effectively identifies deleterious variants in key non-coding elements such as transcription factor binding sites, enhancers, and DNase I hypersensitive sites, focusing on regions under high selective pressure and their impact on gene regulation. Using integrated data from large-scale resources (including ENCODE and 1000 Genomes Project) with cancer genomics data, FunSeq can prioritize known TERT promoter driver mutations. Building on the success of FunSeq, we have developed **FunSeq2**, an enhanced version that delves deeper into the relationships between noncoding regulatory elements and their associated protein-coding genes [**Fig 6**] [PMID: 25273974]. Utilizing an entropy-based scoring system, FunSeq2 evaluates the correlation between enhancer and promoter activity across various ENCODE cell-lines and tissues, facilitating the identification of critical links between regulatory elements and target genes [PMID: 24092746]. This advanced tool is particularly effective at discerning how a single regulatory variant may influence multiple genes, either through direct regulation or because the gene itself serves as a regulatory factor.

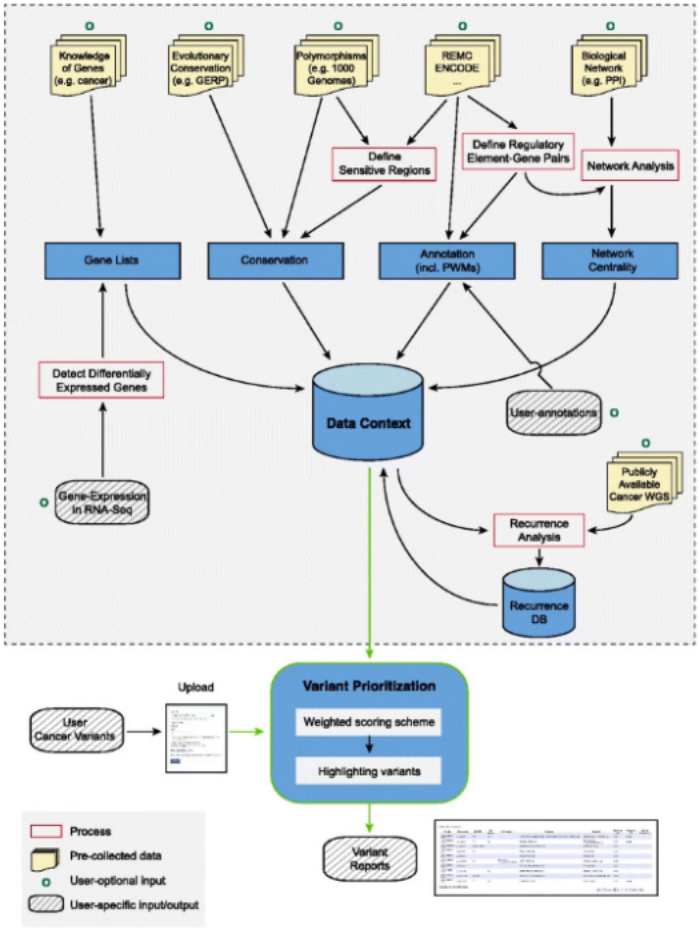

Fig. 6: Funseq2 workflow and data context.

Rare somatic and germline burden tests. We have refined statistical methods for analyzing non-coding regulatory regions with tools like **LARVA** (Large-scale Analysis of Recurrent Variants in noncoding Annotations) and **MOAT** (Mutations Overburdening Annotations Tool). LARVA identifies significant mutation enrichments in both non-coding and coding elements by comparing observed mutation counts with expected counts, adjusted for whole-genome background mutation rates and DNA replication timing biases [PMID: 26304545]. This tool has been effectively used in a pan-cancer analysis of 760 cancer genomes, accurately recapitulating known cancer drivers such as the TERT and TP53 promoters. On the other hand, MOAT offers an alternative empirical approach to mutation burden analysis, utilizing permutations of the input data to evaluate mutation enrichments [PMID: 29121169]. This method supports both annotation-based and variant-based permutations, providing a robust framework for understanding the genetic underpinnings of cancer and other diseases.

Allelic analysis. Allele-specific variants (ASVs) have the potential to provide a highly direct readout of the functional impact of a variant, as they are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE). In our **EN-TEx** project, we created the largest catalog of non-coding AS variants available [PMID: 37001506] (**Fig. 7**). Leveraging this catalog, we developed a deep-learning transformer model, specifically DNABERT, to predict the allele-specific activity of

variants based only on the local nucleotide-sequence context, with a focus on transcription-factor-binding motifs particularly sensitive to variants. This model is designed to determine whether a single-nucleotide variant (SNV) would exhibit AS activity for CTCF binding (as well as other regulatory marks such as POLR2A and various histone modifications), by analyzing the sequence within a 250-bp window around the SNV. This approach highlights the significance of sequence context, indicating that the specific arrangement of nucleotides can greatly influence gene regulatory interactions. The transformer model employs "attention" mechanisms that concentrate on specific sequence positions, often corresponding to known motifs, which allows the model to pinpoint areas within the genome that are most susceptible to these variants.

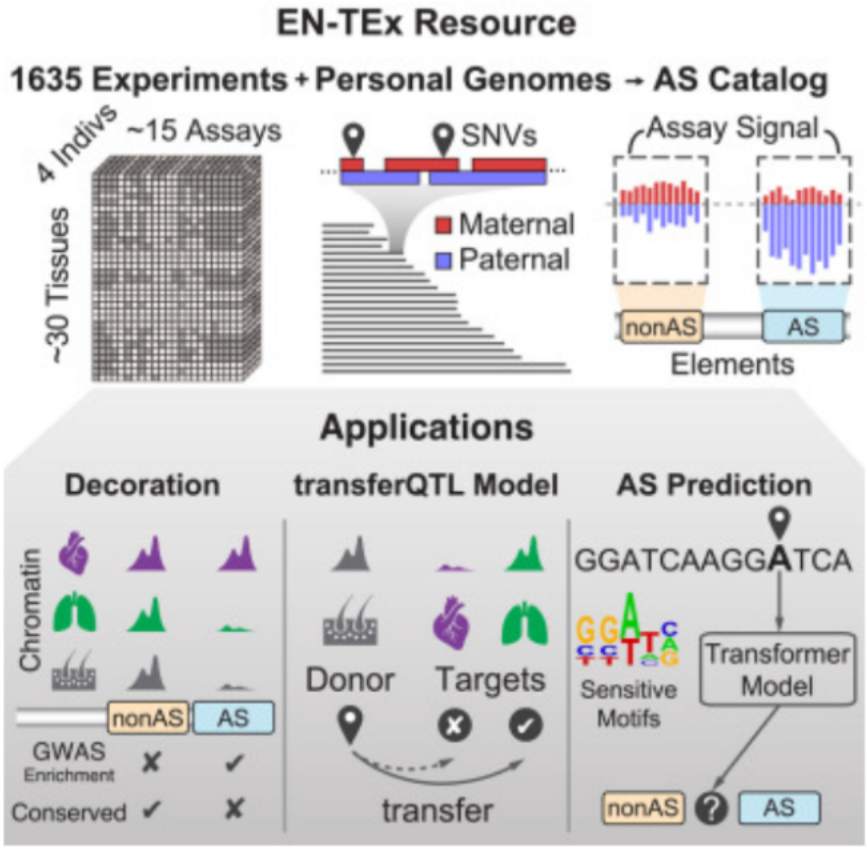

Fig. 7: EN-TEx resource for AS catalog and LLM model for AS prediction.

In addition to the EN-TEx resources and the LLM model, we have developed tools like **AlleleSeq** and **AlleleDB** to further investigate allele-specific phenomena [PMID: 21811232, 27089393]. AlleleSeq helps detect candidate variants associated with allele-specific binding (ASB) and expression (ASE), while AlleleDB provides a comprehensive analysis of allele-specific activity across diverse RNA-Seq and ChIP-seq data sets, offering insights into the distribution and impact of ASVs [PMID: 24037378, 22955616, 23128226]. These tools, supported by robust statistical methods, allow for the detailed examination of allelic variation and its regulatory effects, solidifying our understanding of AS behavior in genomic research.

Summary and future directions. In summary, our laboratory has developed a robust suite of tools and databases designed to tackle various facets of genomic analysis, encompassing the study of coding and non-coding variants, rare somatic and germline burden tests, and detailed allelic analysis. These resources, including ALoFT for annotating loss-of-function variants,

FunSeq and FunSeq2 for analyzing non-coding regions, LARVA for identifying mutation enrichments, AlleleSeq for detecting allele-specific variants, and the EN-TEx project utilizing deep-learning transformer models (LLMs) for predicting allele-specific activity, provide a comprehensive foundation for exploring complex genomic datasets. To contribute effectively to the SysBio platform, we plan to integrate and expand these tools to connect our genomic database with additional data types such as clinical outcomes, phenotypic data, and patient-reported outcomes. This integration will allow for cross-disease analysis and the identification of shared molecular mechanisms across various tissues. Potential deliverables for the SysBio platform include a unified data integration tool that combines the capabilities of our existing tools into a single platform, enhanced visualization tools for complex data interaction, and a user-friendly interface designed for efficient navigation and manipulation of integrated datasets. These enhancements will not only augment our current genomic analysis and database capabilities but also support the broader goals of the SysBio initiative, facilitating advanced research in systems biology and personalized medicine.

5. **General Tool Development**

The Gerstein laboratory played a large role in the development of DOE Systems Biology Knowledgebase (KBase) [PMID: 29979655]. KBase is an open-source platform facilitating the integration of diverse biological data for predictive modeling, with a focus on microbes, plants, and their communities. KBase consolidates extensive genomic, metabolic, and environmental data from various repositories, offering users a web-based interface for data sharing, analysis, and visualization. Key features include comprehensive support for data analysis and reproducibility, flexible data sharing, and a user-friendly point-and-click interface, alongside over 160 built-in applications covering tasks such as genome assembly, metabolic modeling, and RNA-seq analysis. The platform's data model ensures interoperability and encourages community contributions through its Software Development Kit. KBase has garnered significant adoption, with thousands of users creating reproducible computational workflows called Narratives (**Figure 8**), leading to published research on topics like microbial metabolic modeling and trophic interactions within microbial communities. Future development aims to enhance collaboration and knowledge discovery within the scientific community.

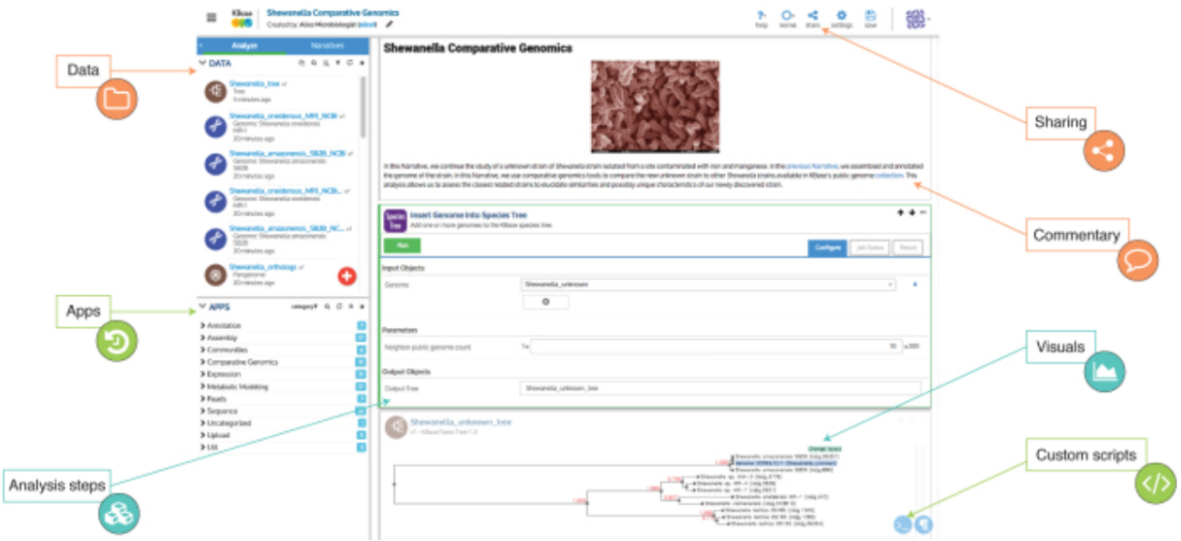

Fig 8. Example of a Narrative in KBase and its respective features.

In the field of genomics, accessing and analyzing vast amounts of data efficiently and effectively remains a critical challenge. Traditional methods of interaction with genomic databases, often involving manual input via mouse and keyboard, are cumbersome and impractical in various settings, such as during presentations or in laboratories where researchers may wear gloves. Recognizing this gap, Gerstein Lab developed "Gene Tracer," an innovative voice-controlled tool designed to enhance the interactive querying and visualization of genomic information. This tool is built on multi-cloud environments, primarily utilizing Amazon's cloud services to handle the backend computational needs and data storage effectively. By deploying the backend server on Amazon EC2 and employing other Amazon Web Services (AWS) like Lambda for server-side operations, the lab ensures scalability and reliability of Gene Tracer. This cloud-based approach not only meets the computational demands of processing large genomic datasets but also guarantees that the system remains responsive and accessible to users from any location. This adaptation to technological advances in cloud computing helps address the specific needs of genomic research and showcases a shift towards multi-cloud strategies to optimize resource use and data management in scientific computing [PMID: 33792640].

Traditional programming approaches often struggle to keep pace with the specificity and volume of data generated in bioinformatics. Recognizing this need, Prof. Mark Gerstein's lab has developed BIOCODER, a benchmark for evaluating the ability of large language models (LLMs) to generate bioinformatics-specific code. This initiative highlights Gerstein Lab’s expertise in integrating APIs and microservices to enhance bioinformatics data analysis. BIOCODER is designed to challenge and assess the performance of LLMs across a spectrum of bioinformatics coding tasks, such as managing diverse biological data formats and processing workflows. It emphasizes the use of APIs from various packages, reflecting real-world scenarios where bioinformaticians must integrate multiple data sources and software tools. The benchmark includes over two thousand bioinformatics-specific coding problems and employs a robust testing framework using Docker to ensure that evaluations mimic practical use cases. This approach not only advances the field by providing a specialized tool for improving the performance of code-generating models but also showcases Prof. Gerstein's pioneering work in leveraging modern

software architecture principles, such as microservices, to address complex challenges in bioinformatics [cite: doi.org/10.48550/arXiv.2308.16458].

6. **Biological networks and analysis**

The complexity of biological systems and the explosion of genomic data have driven an urgent need for sophisticated tools to analyze and interpret these vast datasets. Understanding the intricacies of genetic networks, protein interactions, and cellular pathways is fundamental for advancements in medicine and biology. Professor Mark Gerstein's work addresses this need by developing comprehensive bioinformatics tools that leverage network theory to elucidate the relationships and dynamics within various biological systems.

In the paper "Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data," Professor Mark Gerstein and his colleagues develop a comprehensive framework for understanding the complex interactions within gene regulatory networks by integrating high-throughput sequencing data across various levels of regulation. The study presents an integrated regulatory network (IRN) that encompasses transcription factors (TFs), microRNAs (miRNAs), and their respective gene targets. This network reveals the complex interplay between different types of regulatory molecules, providing insights into both transcriptional and post-transcriptional regulation [PMID: 22125477]. The integrated network combines data from ChIP-Seq, which identifies TF binding sites, and RNA-Seq, which profiles expression levels, creating a multifaceted view of gene regulation. The paper emphasizes the novel use of this integrated approach to map regulatory interactions in a system-wide manner, marking a significant advance over previous studies that typically focused on one regulatory level at a time. By doing this, Gerstein's team was able to uncover new insights into how TFs and miRNAs co-regulate genes and participate in complex regulatory circuits that maintain cellular functions [PMID: 22125477]. The integrated network combines data from ChIP-Seq, which identifies TF binding sites, and RNA-Seq, which profiles expression levels, creating a multifaceted view of gene regulation. The paper emphasizes the novel use of this integrated approach to map regulatory interactions in a system-wide manner, marking a significant advance over previous studies that typically focused on one regulatory level at a time. By doing this, Gerstein's team was able to uncover new insights into how TFs and miRNAs co-regulate genes and participate in complex regulatory circuits that maintain cellular functions[PMID: 22125477].

In another work, Professor Mark Gerstein and his colleagues address the challenge of understanding how microbial communities adapt to diverse environmental conditions through a network-based analytical approach, primarily using metagenomic data. Recognizing the limitations of traditional studies that classified environments in overly broad categories (e.g., terrestrial vs. marine), they advanced the field by quantifying the correlation between environmental factors and microbial metabolic pathways. The innovative approach involved the use of canonical correlation analysis (CCA) and discriminative partition matching (DPM) to link specific environmental features with microbial metabolic activities[PMID: 19164758] . Their study went beyond simple correlation by employing a sophisticated multivariate analysis that allowed them to explore the complex interplay between numerous environmental variables and various levels of metabolic processes. This method facilitated a deeper understanding of how microbial communities adapt their metabolic pathways to their surroundings, providing a detailed picture of environmental impacts on microbial ecology. The novel concept of a "metabolic footprint," which identifies the ensemble of metabolic pathways that correlate with environmental conditions, was introduced. This concept allows researchers to predict environmental conditions based on

microbial metabolic profiles, offering potential applications in environmental monitoring and assessment [PMID: 19164758].

In another work, Professor Mark Gerstein and his team developed advanced computational tools to analyze the human regulatory network using ENCODE project data. They tackled the complex task of understanding how transcription factors (TFs), which regulate gene expressions, interact within a vast network to control genetic activities across different cells and conditions. The research focused on detailing the architecture of this network by identifying and mapping the genomic binding sites of 119 TFs across various cell types. The team utilized ChIP-Seq data to establish where these TFs bind in the genome, leading to the discovery of combinatorial and highly context-specific associations between different TFs. These findings are critical as they show that TF binding is not random but rather highly orchestrated, with specific TF combinations binding at unique locations to execute cellular functions. Gerstein's group organized this complex interaction data into a hierarchical structure that revealed new insights into the regulatory landscape of human genes. They integrated other types of genomic data, such as microRNA (miRNA) interactions and protein interactions, creating a comprehensive meta-network that illustrates how different levels of regulation interact to influence gene expression [PMID: 22955619].

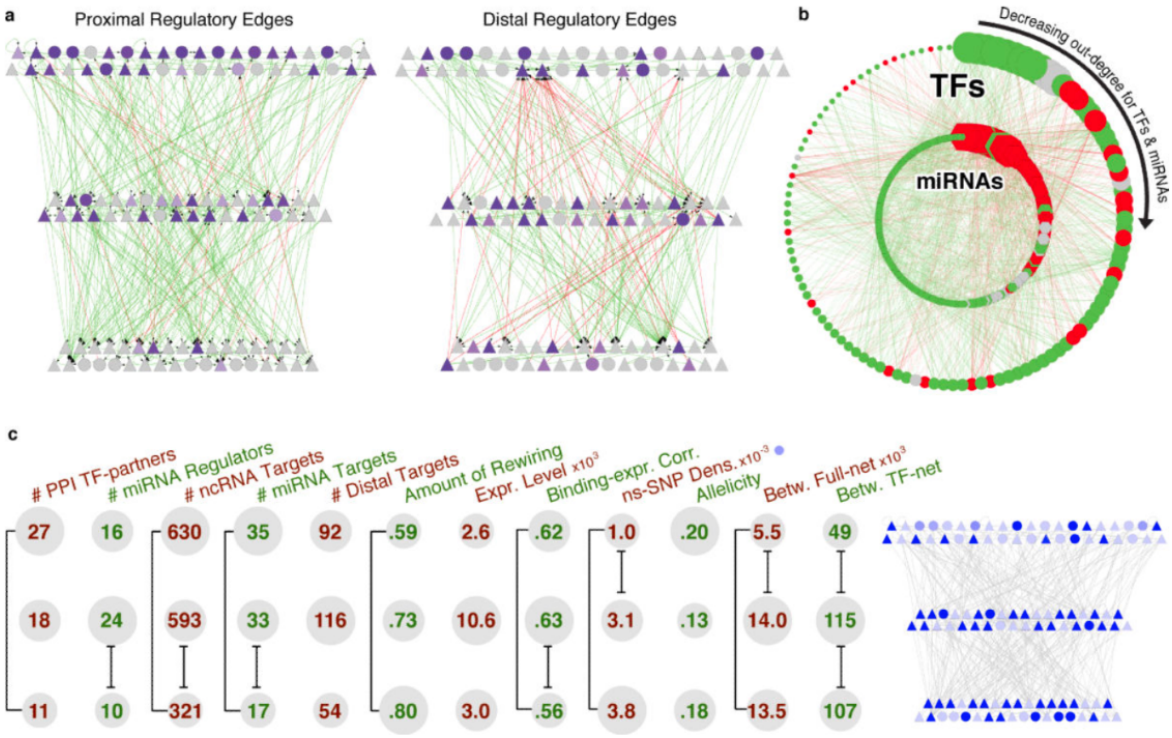

Figure 9. (a) Close-up of the TF hierarchy. (b) Close-up of the TF-miRNA regulation. (c) Average values of various properties for each level are shown for the proximal-edge hierarchy.

In the paper titled "Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights," Professor Mark Gerstein and his team pioneer an innovative approach to enhancing the understanding of protein interaction networks through the integration of structural biology with network analysis. They meticulously developed a method to annotate interactions within a consensus yeast interaction network using atomic-resolution data from three-dimensional protein structures. This led to the creation of a Structural Interaction Network (SIN), which

classifies interactions as either mutually exclusive or simultaneously possible, based on whether interaction partners bind to the same or different interfaces on a common protein. This methodological innovation provides a more nuanced view of protein networks, distinguishing between different types of protein interactions that were previously undifferentiated. By analyzing these interactions in terms of their structural and functional implications, Gerstein's team was able to derive deeper insights into the dynamics of protein networks. For instance, they discovered that proteins connected by simultaneously possible interactions often share similar functions and are co-expressed, suggesting stable, possibly permanent, associations. Conversely, mutually exclusive interactions are indicative of more transient interactions within the network. Gerstein's research elucidates the physical constraints on protein interactions, highlighting how structural properties can influence biological function and evolution. This work significantly advances our understanding of how protein structures shape the behavior of biological networks and provides a valuable framework for predicting protein interactions based on structural data [PMID: 17185604].

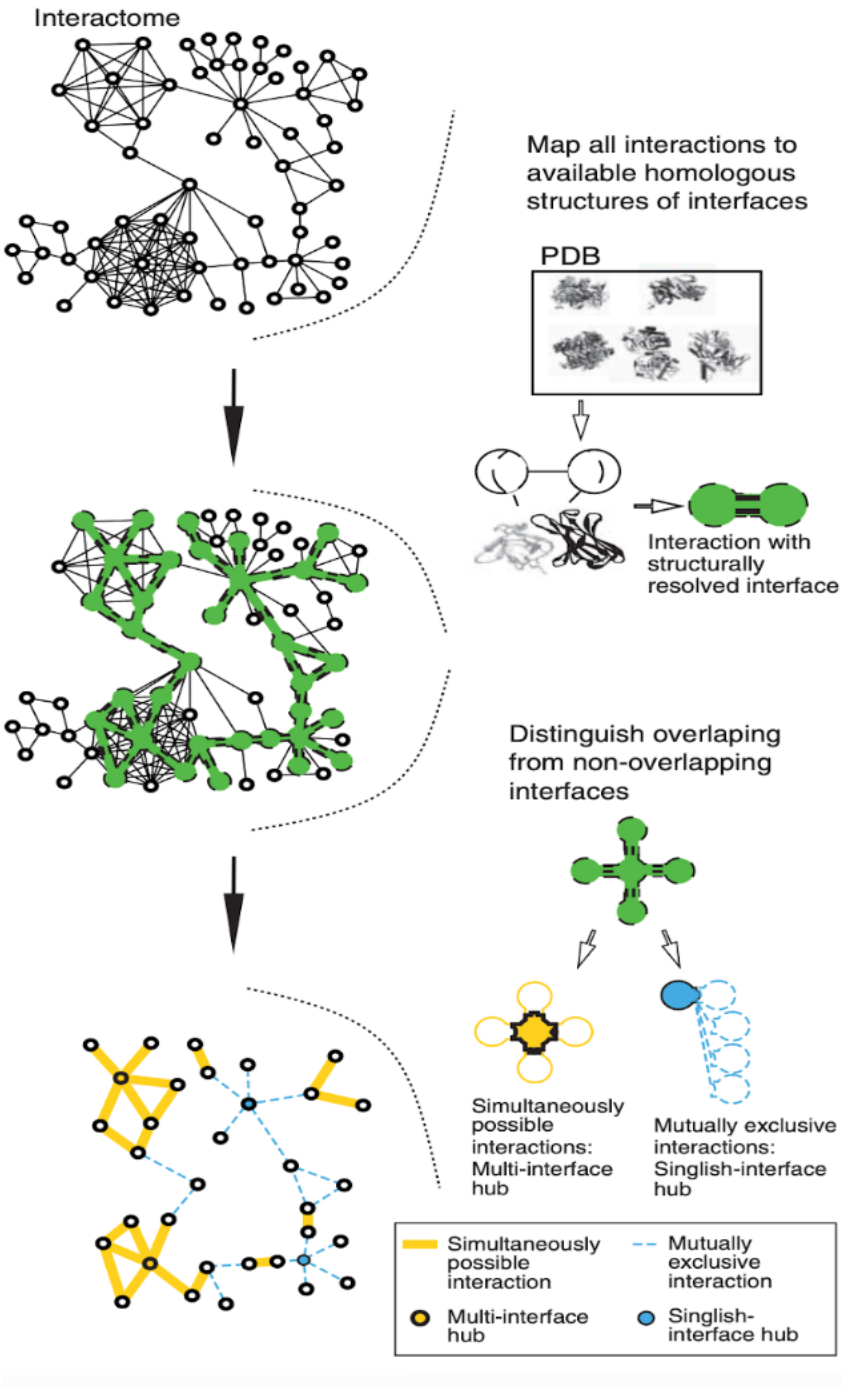

Figure 10. The creation of a structural interaction network (SIN) dataset.

This paper addresses the problem of understanding protein-protein interaction networks (interactomes) by incorporating structural modeling with network analysis. To embed proteins in a graph, interactions from a filtered protein interaction dataset are mapped to Pfam domains. Pfam is a database of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). These Pfam domains are then linked to known structures of protein interactions using iPfam, a database of known interactions between Pfam domains in the Protein Data Bank (PDB). Interactions where both interacting partners and their homologous

domains can be found in a 3D structure of a protein complex are kept. The interactions are then categorized as mutually exclusive or simultaneously possible based on 3D structural exclusion, which considers the binding interfaces used by proteins interacting with a common partner.

To build a SIN network, the study then examines properties of proteins linked by the two types of interactions defined. Proteins connected by simultaneously possible interactions are more likely to share the same function and be coexpressed than those connected by mutually exclusive interactions. The multi-interface hubs (proteins with several interaction interfaces) are more likely to be essential and coexpressed with their interaction partners, offering a structural basis for the existence of different types of expression dynamics in hubs. The paper posits that this distinction has implications for understanding the mechanisms of network evolution.

By using machine learning models that understand both the network topology (from PPI data) and the structural constraints (from 3D data), predictions about unknown protein interactions can be improved. The study found that proteins connected by simultaneously possible interactions tend to share functions and expression patterns. This suggests that the SIN can provide functional insights that pure PPI networks may not reveal. Multi-interface hubs—proteins that can bind to several others simultaneously—are more likely to be essential and evolutionarily conserved than singlish-interface hubs. This structural perspective adds a layer of understanding to network centrality, beyond what is inferred from topological data alone.

In conclusion, the approach of creating a SIN that incorporates both PPI data and 3D structural data provides a richer and more nuanced understanding of protein interactions within a cell. It distinguishes between the types of interactions based on physical structural constraints, thus refining the understanding of hub proteins in cellular processes. This method also paves the way for more accurate machine learning models that can predict protein interactions and functions within the cell, leveraging the additional structural information to enhance their accuracy [PMID: 17185604].

Gerstein's lab also has extensive experiences in utilizing multi-modalities. In the EN-TEx study, the use of multi-modalities is exemplified by the integration of a large-scale dataset combining genomics, transcriptomics, and epigenetics from multiple tissues. By mapping over 1,635 datasets to four personal genomes across roughly 30 tissues with about 15 assays each, the study created a comprehensive catalog of allele-specific activity. This multi-layered approach allowed for the nuanced observation of regulatory elements, offering insights into the allele-specific binding and expression that could be correlated with genetic variants and impacts on health and disease [PMID: 37001506].

Another instance of utilizing multi-modalities in the EN-TEx project is seen in the development of predictive models for the impact of genetic variants. The researchers harnessed the richness of their multi-assay data to inform a transformer model capable of predicting allelic activity based on local sequence context. This reflects the powerful synergy that can be achieved when different types of biological data are combined to provide a more holistic view of genomic function. The cross-tissue, cross-individual, and cross-assay aggregation strategies enhanced the detection power of allele-specific events, thus enabling the generation of a sizable catalog of such events that can be used to predict variant impact with high accuracy [PMID: 37001506].