BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Dai, Yulin

eRA COMMONS USER NAME (credential, e.g., agency login): DAI_YULIN

POSITION TITLE: Research Assistant Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Start Date MM/YYYY	Completion Date MM/YYYY	FIELD OF STUDY
University of Science and Technology of China, Hefei, Anhui, China	B.Sc.	09/2006	07/2010	Systems Biology
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai	Ph.D.	09/2010	07/2015	Bioinformatics
Vanderbilt University Medical Center, Nashville, TN	Postdoctoral Fellow	02/2016	07/2017	Bioinformatics
University of Texas Health Science Center at Houston, Houston, TX	Postdoctoral Fellow	11/2017	12/2019	Bioinformatics

A. Personal Statement

I have interdisciplinary training and nearly 10 years of bioinformatics and statistical genetics research experience. My PhD work focused on application of the evolution theory to solve problems in population genetics/genomics. I also received solid training in constructing bioinformatics pipelines and utilizing statistical methods to analyze different types of next-generation sequencing (NGS) data (e.g. whole genome sequencing, whole exome sequencing, RNA-sequencing, ChIP-seq, single-cell RNA-seq/ATAC-seq) in many collaboration-based projects. In my postdoc training at Vanderbilt University Medical Center, I joined the Undiagnosed Disease Network (UDN) as a bioinformatician to analyze the NGS data generated from case-parent trio study. I built a tool to prioritize these latent pathogenic genes based on the phenotypes of patients. After I joined the School of Biomedical Informatics at UTHealth in 2017, **I focused on methodology development and application for detecting or inferring causal variants of complex diseases related to human brain from the following aspects**. 1) I developed integrative frameworks to harness the genetic risks with the top odd ratio among multimodal data. My methods have been applied to multiple complex traits (including psychiatric disorders, and multiple sclerosis) using various omics data types, such as genetic (e.g., genome-wide association studies (GWAS), rare variants, *de novo* mutations, and copy number variants), transcriptomics (bulk RNA-seq and scRNA-seq) and epigenetic (e.g., DNA methylation and ChIP-seq) data. 2) I also conducted cutting-edge statistical methods (colocalization and TWAS) to fine-map the noncoding loci in COVID-19 severe symptoms and amyotrophic lateral sclerosis. 3) I utilized network-diffusion-based approaches to connect the genetic implications (individual or population) and epi/genomics alternatives on protein-protein-interaction (PPI) to prioritize gene modules that potentially could be targeted for drug repurposing. These approaches have applied to Alzheimer's Disease, psychiatric disorders and multiple sclerosis. 4) I developed tools, databases, and web servers to systematically assess the tissue- and cell-type context of genesets (e.g. complex traits and disease). **Overall, I had the training and expertise to analyze various high-throughput epi/genomics data and develop methodologies to connect the genetic basis of human complex diseases to their disease context and eventually potential treatment targets.**

Ongoing and recently completed projects that I would like to highlight include:

1U01AG079847 Zhao, Jiang (PI) Role: Co-I 04/01/2023-3/31/2028

AIM-AI: an Actionable, Integrated and Multiscale genetic map of Alzheimer's disease via deep learning
The major goals are to develop and implement a robust artificial intelligence (AI) framework, namely AIM-AI, for transforming the genetic catalog of Alzheimer's Disease in a way that is Actionable, Integrated and Multiscale, so that genetic factors have clear utility for subsequent studies focusing on cognitive systems.

R01DE030122-01A1 Zhao (PI) Role: Co-I 04/01/2021-3/31/2024
Deep learning for decoding genetic regulation and cellular maps in craniofacial development
We combine statistical and bioinformatics analytical strategies to obtain the disease risk of common genetic variants and *de novo* mutations through human craniofacial development.

R01LM012806-05 Zhao (PI) Role: Co-I 09/01/2021-05/31/2025
Predicting Phenotype by Deep Learning Heterogeneous Multi-Omics Data
To develop a deep learning method for variant impact predictor, that maximally utilizes functional and regulatory data to predict the causal roles of variants in complex diseases and traits in cell type-specific and developmental stage-specific manner.

Citations:

a. **Dai Y**, Jia P, Zhao Z, Gottlieb, A (2022). A Method for Bridging Population-Specific Genotypes to Detect Gene Modules Associated with Alzheimer's Disease. *Cells*, 11(14), 2219. PMCID: PMC9319087 *[contributed **AIzheimer's Disease research**]*
b. Liu A, Manuel AM, **Dai Y**, Fernandes BS, Enduru N, Jia P, Zhao Z (2022). Identifying candidate genes and drug targets for Alzheimer's disease by an integrative network approach using genetic and brain region-specific proteomic data. *Human Molecular Genetics*, PMID: 35640139 *[contributed **Alzheimer's Disease research**]*
c. Li X, Fernandes B, Liu A, Lu Y, Chen J, Zhao Z*, **Dai Y*** (2023). Genetically-regulated pathway-polygenic risk score (GRPa-PRS): A risk stratification method to identify genetically regulated pathways in polygenic diseases. *medRxiv (***under reviewed in Genome Medicine***)*, 2006.2019.2329162. *[contributed **Alzheimer's Disease resilience research**]*
d. **Dai Y#,** Wang J#, Jeong HH, Chen W, Jia P, Zhao, Z (2021). Association of CXCR6 with COVID-19 severity: delineating the host genetic factors in transcriptomic regulation. *Human Genetics*, 1–16. PMCID: PMC8216591 *[contributed human genetics reserach]*

B. Positions, Scientific Appointments and Honors

Positions and Scientific Appointments

2022 - Present Assistant Director of Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston
2020 - Present Research Assistant Professor, Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
2021 - Present Editorial Board, Frontiers in Systems Biology
2019 - Present Program Committee, International Conference on Intelligent Biology and Medicine 2020-2023

Honors

2021, 2022 American Society of Human Genetics, Reviewer's Choice abstract (Top 10% ranked abstract)
2020 Travel Award, Americas Committee for Treatment and Research in Multiple Sclerosis Young Scientist Summit
2018, 2019 American Society of Human Genetics, Reviewer's Choice abstract (Top 10% ranked abstract)
2018 Travel Award, International Conference on Intelligent Biology and Medicine
2015 Merit Student Award by Shanghai institute for biological sciences

C. Contributions to Science

1. **Multimodalities integration of complex diseases**. In my current school, I collaborated with complex disease specialists and performed genetic and genomic studies of multiple human complex diseases including Crohn's Disease, alcohol use disorder (AUD), opioid use disorder (OUD), and multiple sclerosis (MS). Specifically, I led to develop an integrative analysis using multi-omics data such as GWAS, eQTL, DNA methylation, DNA accessibility, and histone modification with applications in Crohn's disease and AUD as well. With multiple lines of evidence from various omics analyses, we could prioritize a list of genes containing the maximized information linking to diseases, which could help to better understand the etiology of human complex diseases. Next, we generated matched transcriptomic and DNA methylation profiles of 22 OUD subjects and 19 non-psychiatric controls. We developed a novel co-expressed network framework to discover gene modules with coherent regulation within transcriptomic and methylomic levels. Lastly, we refined one Bayesian framework to capture a high-confidence MS gene set with maximized posterior possibility from diverse modalities such as genomic, epigenomic, eQTL, and single-cell omics. Our MS gene set could further serve as a benchmark of MS GWAS risk genes for future validation or genetic studies. Overall, I have been focusing on developing methods to prioritize genetic implications and genomics alternatives that underlie complex diseases and use advanced statistical methods to connect the genetic risk factors and potential druggable targets.

    a. **Dai Y**, Pei G, Zhao Z, Jia P (2019). A convergent study of genetic variants associated with Crohn's Disease: evidence from GWAS, gene expression, methylation, eQTL and TWAS. *Frontiers in Genetics*, 10:318. PMCID: PMC6467075
    b. **Dai Y**, Hu R, Pei G, Zhao Z, Jia P (2020). Diverse types of genomic evidence converge on alcohol use disorder risk genes. *Journal of Medical Genetics,* jmedgenet-2019-106490. PMCID: PMC7487038
    c. Liu A, **Dai Y**, Mendez EF, Hu R, Fries GR, Najera KE, Jiang S, Meyer TD, Stertz L, Jia P and Walss-Bass C (2021). Genome-wide correlation of DNA methylation and gene expression in postmortem brain tissues of opioid use disorder patients. *International Journal of Neuropsychopharmacology*. PMCID: PMC8598308
    d. Liu A, Manuel AM, **Dai Y**, Zhao Z (2022). Prioritization of risk genes in multiple sclerosis by a refined Bayesian framework followed by tissue-specificity and cell type feature assessment. *BMC genomics*, 23(4), 1-17. PMCID: PMC9092676

2. **Statistical method development, GWAS analysis and statistical fine-mapping.** We developed a novel computational framework that could identifiy resilience-related genetically-regulated pathways (GRPa) from individual genetically-regulated expression of Alzheimer's Disease cohort utilizing a unique polygenic risk score (PRS)-based stratification strategy. We successfully identified well-known APOE-related function, including amyloid-beta clearance and tau protein binding. More importantly, we first-time identified a few GRPas related to AD resilience effect, including synapse function and thioester hydrolase activity, which provides new insights into the pathways linked to AD risk and resilience. During the COVID-19 pandemic, we explore the genetic risks underlying COVID-19 severe patients. We used cutting-edge statistical methods to delineate the host genetic factors and pinpointed *CXCR6* as the risk gene in *3p21.31* locus. **This discovery was mentioned in a few medical news and received many positive feedbacks from the research community**. In the second project, we conducted genome-wide association analyses on 640 circulating metabolites in 3,926 Hispanic Community. We performed the multi-traits colocalization analysis and mendelian randomization to fine-map the 46 genome-wide significant loci. In the last project, we conducted comprehensive cutting-edge statistical approaches (transcriptome-wide association study, colocalization, summary-based mendelian randomization) to pinpoint the genes related to Amyotrophic lateral sclerosis GWAS loci. We identified 43 genes at 24 loci, including 23 novel genes and 10 novel loci. In summary, I have adequate experience in both GWAS analysis and downstream fine-mapping analysis utilizing cutting-edge statistical methods.

a. Li X, Fernandes B, Liu A, Lu Y, Chen J, Zhao Z*, **Dai Y*** (2023). Genetically-regulated pathway-polygenic risk score (GRPa-PRS): A risk stratification method to identify genetically regulated pathways in polygenic diseases. *medRxiv* (under reviewed in Genome Medicine), 2006.2019.2329162. [contributed Alzheimer's Disease resilience research]
b. **Dai Y#**, Wang J#, Jeong HH, Chen W, Jia P, Zhao, Z (2021). Association of CXCR6 with COVID-19 severity: delineating the host genetic factors in transcriptomic regulation. *Human Genetics*, 1–16. PMCID: PMC8216591
c. Feofanova EV, Chen H, **Dai Y**, Jia P, Grove ML, Morrison AC, Qi Q, Daviglus M, Cai J, North KE and Laurie CC (2020). A genome-wide association study discovers 46 loci of the human metabolome in the Hispanic community health study/study of Latinos. *The American Journal of Human Genetics*, 107(5), pp.849-863. PMCID: PMC7675000
d. Pan S, Liu X, Liu T, Zhao Z, **Dai Y,** Wang YY, Jia P, Liu F (2022). Causal Inference of Genetic Variants and Genes in Amyotrophic Lateral Sclerosis. *Frontiers in Genetics* 13: 917142. PMCID: PMC9257137

3. **Computational tools, databases and web server.** My third research interest is to uncover the tissues and cell types context of complex traits and diseases. I have led the development of a series of computational tools, databases and web servers to decode tissue- and cell-type- specificity based on gene expression profiles. We constructed the Tissue-Specific Enrichment Analysis DataBase (TSEA-DB) for >5000 GWAS datasets with inferred causal tissues using deTS methods we previously developed. We identified thousands of significant trait-tissue associations, which both recapitulate known biology and provide new insights into the pathogenesis of human phenotypes. With the advances in single-cell technology, scientists have been empowered to explore transcriptomics at single-cell resolution. I led the cell type-specific enrichment analysis (CSEA) of thousands of GWAS datasets across ~1355 tissue cell types from 110 single-cell transcriptome panels across 11 human organ systems as an omnibus collection for systematic exploration of human complex trait and cell type associations (CSEA-DB). Lastly, to benefit the large research and clinical community, we built the first web-based application to quickly assess the biological context of genes (WebCSEA). Overall, I have extensive experience in developing computational tools to contextualize the tissues and cell types of complex traits and diseases. We also developed several databases and web servers to store and visualize such associations.

   a. Pei G, **Dai Y**, Zhao Z, Jia P (2019) deTS: tissue-specific enrichment analysis to decode tissue specificity. *Bioinformatics* 35(19):3842-3845. PMCID: PMC6761978
   b. Jia P#, **Dai Y# [co-first author]**, Hu R, Pei G, Manuel AM, Zhao Z (2019). TSEA-DB: a trait–tissue association map for human complex traits and diseases. *Nucleic Acids Research*, 48(D1):D1022-D1030. PMCID: PMC7145616
   c. **Dai Y#**, Hu R#, Manuel AM, Liu A, Jia P, Zhao Z (2020). CSEA-DB: An omnibus for human complex trait and cell type associations. *Nucleic Acids Research*. 49.D1: D862-D870. PMCID: PMC7778923
   d. **Dai Y#**, Hu R#, Liu A, Cho KS, Manuel AM, Li X, Dong X, Jia P, Zhao Z (2022). WebCSEA: Web-based Cell-type Specific Enrichment Analysis of Genes. *Nucleic Acids Research*, *gkac392*. PMID: 35610053

4. **Network-based methods.** In my current lab, I have conducted many projects that utilize network-diffusion approaches to link gene-level modalities to co-expression networks or the human protein-protein interactions (PPI) reference. In the first project, we systematically characterized the spatial-temporal changes of brain development and aging co-expression network, we projected the genetic impact onto the dynamic network and identified several key time points and brain regions that enriched with psychiatric disorder genes activities. AD impacts all ethnicities, given the slight difference in genetic basis. We want to understand whether AD patients with different ethnic backgrounds would share their genetic implications on certain modules of human reference PPI. Interestingly, we discovered ~180 AD-associated genes from gene modules shared between Whites, African descent and Hispanics populations, highlighting new gene modules associated with AD. In the third work, we applied our in-house network-based tool, Edge-Weighted Dense Module Search of GWAS (EW_dmGWAS) to assess the module genes with combined genetic risk and proteome alterations in brain region-specific manner. We highlighted a few molecular pathways underlying AD pathogenesis and explore their potential drug targets. Lastly, we developed a novel methodology (**scGWAS**) to expand our dense module search algorithm to accommodate single-cell RNA-seq data, we compare the genetic enrichment of more than 40 human complex traits and diseases among 18 human tissues. We showed that the trait-cell type associations identified by scGWAS, while generally

constrained to trait-tissue associations. Overall, I have extensive experience to utilize network-diffusion-based approaches to connect the individual or population genetic implications with other modalities onto protein-protein-interaction (PPI) to prioritize gene modules that potentially could be targeted for drug repurposing.

a. **Dai Y**, Jia P, Zhao Z and Gottlieb A (2022). A Method for Bridging Population-Specific Genotypes to Detect Gene Modules Associated with Alzheimer's Disease. *Cells*, 11(14), p.2219. PMCID: PMC9319087 [contributed Alzheimer's Disease research]
b. Liu A, Manuel AM, **Dai Y**, Fernandes BS, Enduru N, Jia P, Zhao Z (2022). Identifying candidate genes and drug targets for Alzheimer's disease by an integrative network approach using genetic and brain region-specific proteomic data. *Human Molecular Genetics*, PMID: 35640139 [contributed Alzheimer's Disease research]
c. **Dai Y**, O'Brien T, Pei G, Zhao Z**,** Jia P (2020). Characterization of genome-wide association study data reveals spatiotemporal heterogeneity of mental disorders. *BMC Medical Genomics* 13.11: 1-14. PMCID: PMC7771094
d. Jia P, Hu R, Yan F, **Dai Y**, Zhao Z. scGWAS: landscape of trait-cell type associations by integrating single-cell transcriptomics-wide and genome-wide association studies. *Genome biology* 23, no. 1 (2022): 1-24.,. PMID: 36253801 [contributed single-cell network-based approach]

5. **Machine learning collaboration.** Lastly, I have accumulated adequate knowledge and experience in collaborating with machine learning and deep learning experts in various fields such as noncoding variants function prediction, drug repurposing, and gene manifolding. Since ~90% of genetic variants from GWAS are located in non-coding regions, we implemented a deep-learning-based convolutional neural network (CNN) model (DeepFUN) to predict the regulatory roles of genetic variants across ~8000 epigenomic modifications profilings collected from ENCODE and Roadmap consortia. Drug response differs substantially in cancer patients due to inter- and intra-tumor heterogeneity, we developed a deep variational autoencoder (VAE) model to compress thousands of genes into latent vectors in a low-dimensional space. We then demonstrated that these encoded vectors could accurately impute drug response, outperform standard signature-gene-based approaches, and appropriately control the overfitting problem. The famous Word2vec transforms each word into a high-dimensional vector and learns the latent correlation between words in those embeddings. We developed a method Gene2vec to project all human genes to high-dimensional embeddings to predict their function using co-expression profiles. In summary, I have been involved in large-scale machine learning and deep learning projects, which will assist me to apply machine learning approaches to fulfill projects utilizing machine learning and deep learning techniques.
   a. Pei G, Hu R, **Dai Y**, Manuel AM, Zhao Z and Jia P (2021). Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations. *Nucleic acids research*, 49(1), pp.53-66. PMCID: PMC7797043
   b. Jia P, Hu R, Pei G, **Dai Y**, Wang YY and Zhao Z (2021). Deep generative neural network for accurate drug response imputation. *Nature communications*, 12(1), pp.1-16. PMCID: PMC9257137
   c. Du J, Jia P, **Dai Y**, Tao C, Zhao Z and Zhi D (2019). Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(1), pp.7-15. PMCID: PMC6360648

Complete List of Published Work in My Bibliography:
https://www.ncbi.nlm.nih.gov/myncbi/16wyJJh89vtYU5/bibliography/public/