ICIBM

The 13th International Conference on Intelligent Biology and Medicine (ICIBM 2025)

August 3rd - 5th, 2025
Columbus, Ohio, USA

Hosted by:
The International Association for Intelligent Biology and Medicine (IAIBM),
Department of Biomedical Informatics, The Ohio State University, and
Translational Data Analytics Institute, The Ohio State University

Table of Contents

Welcome to ICIBM 2025!

On behalf of all our conference committees and organizers, we are thrilled to welcome you to the 2025 International Conference on Intelligent Biology and Medicine (ICIBM 2025). ICIBM is the official conference of The International Association for Intelligent Biology and Medicine (IAIBM, http://iaibm.org/), a non-profit organization dedicated to advancing intelligent biology and medical science through member collaboration, education, and global networking.

As we step into 2025, the fields of bioinformatics, systems biology, and intelligent computing continue to experience rapid advancements, profoundly influencing scientific research and medical innovations. Building on the successes of previous years, ICIBM 2025 is designed to be a platform for interdisciplinary research, dynamic discussions, educational growth, and collaborative opportunities across these evolving fields.

This year, we are excited to present an exceptional line-up of keynote speakers, including Drs. Veera Baladandayuthapani, Jiang Bian, Qing Nie, and Julie A. Johnson. Additionally, we are honored to feature four eminent scholar speakers: Drs. Kaifu Chen, Yu-Ping Wang, Xiang Zhou, and Yufeng Shen. These distinguished researchers are global leaders in their respective fields, and we are privileged to have them share their insights at ICIBM 2025. The conference will also include twelve workshops, along with presentations from faculty members, postdoctoral fellows, Ph.D. students, and trainee-level awardees, selected from outstanding manuscripts and abstracts. These presentations will highlight the innovative technologies and approaches that define our interdisciplinary fields.

We anticipate that this year's program will be invaluable to advancing research, education, and innovation, and we hope you share our enthusiasm for the exciting opportunities ICIBM 2025 will offer. We would like to express our deepest gratitude to our sponsors, whose generous support has made this event possible. Our sponsors include Admera Health, Complete Genomics, Yeasen Biotechnology Co., Ltd, Singleron Biotechnologies Inc., USA, and Olink.

Finally, our heartfelt thanks go to all members of the ICIBM 2025 committees and our volunteers for their dedication and hard work. Their commitment to making ICIBM 2025 a success is a testament to the strength and resilience of our community.

We hope that the program we have prepared will be thought-provoking, foster collaboration and innovation, and provide an enjoyable experience for all attendees. Thank you for joining us at ICIBM 2025. We look forward to your active participation in all that the conference has to offer!

Sincerely,

Lianbo Yu, PhD
Program Co-Chair,
Associate Professor,
The Ohio State University

Leng Han, PhD
Program Co-Chair, Professor,
Indiana University

Maciej Pietrzak, PhD
Publication Co-Chair, The Ohio State University

Xinghua Shi, PhD
Pulication Co-Chair
Temple University

Lang Li, PhD
General Co-Chair
Professor,
The Ohio State University

Zhongming Zhao, PhD
General Co-Chair
Professor
University of Texas Health Science Center at Houston

ACKNOWLEDGEMENTS

General Chairs
Lang Li, The Ohio State University
Zhongming Zhao, University of Texas Health Science Center at Houston
Qin Ma, The Ohio State University

Program Committee
Lianbo Yu, The Ohio State University
Leng Han, Indiana University
Oindrila Bhattacharyya, The Ohio State University
Weidan Cao, The Ohio State University
Sapuni Chandrasena, The Ohio State University
Xiao Chang, CHOP
Lijun Cheng, The Ohio State University
Mohamed Elsaid, The Ohio State University
Rejuan Haque, The Ohio State University
Matthew Hayes, Xavier University of Louisiana
Tao Huang, Shanghai Institute of Nutrition and Health
Weichun Huang, EPA
Peilin Jia, University of Texas Health Science Center at Houston
Garrett Kinnebrew, The Ohio State University
Jiaying Lai, Johns Hopkins University
Aimin Li, Xi'an University of Technology
Fuhai Li, Washington University at St. Louis
Xiaoming Liu, University of South Florida
Tianle Ma, Oakland University
Joseph McElroy, The Ohio State University
Xiaokui Mo, The Ohio State University
Jiang Qian, Johns Hopkins University
Michal Seweryn, University of Lodz
Rama Shankar, Michigan State University
Li Shen, University of Pennsylvania
Yang Shen, Texas A&M University
Jiao Sun, University of Central Florida
Shulan Tian, May Clinic
Manabu Torii, Kaiser Permanente
Alper Uzun, Brown University
Ece Uzun, Brown University
Jiayin Wang, Xi'an Jiaotong University
Junbai Wang, Radium Hospital
Qing Wang, Univerity of Florida
Junfeng Xia, Institute of Physical Science and Information Technology, Anhui University
Min Xu, Carnegie Mellon University
Jianhua Xuan, Virginia Tech
Yu Xue, Huazhong University of Science and Technology
Huihuang Yan, Mayo Clinic

Rui Yin, University of Florida
Shaojie Zhang, University of Central Florida
Wei Zhang, University of Central Florida
Jim Zheng, University of Texas Health Science Center at Houston

Publication Committee
Maciej Pietrzak, The Ohio State University
Xinghua (Mindy) Shi, Temple University

Workshop/Tutorial Committee
Hongbo Liu, University of Rochester
Yusi Fu, Texas A&M University
Qianqian Song, University of Florida
Pengyue Zhang, Indiana University
Travis Johnson, Indiana University
Chi Zhang, Oregon Health & Science University
Xiang Gao, Loyola University Chicago

Award Committee
Fuhai Li, Washington University in St. Louis

Trainee Committee
Chi Zhang, Oregon Health & Science University
Jingwen Yan, Indiana University

Publicity Committee
Shibiao Wan, University of Nebraska Medical Center
Alper Uzun, Brown University

Local Committee
Joseph McElroy, The Ohio State University
Gang Peng, Indiana University

Schedule

The International Conference on Intelligent Biology and Medicine (ICIBM 2025) Program, August 3-5, 2025
Pomerene Hall, 1760 Neil Ave, Columbus, OH

Sunday, August 3rd, 2025

7:30 AM - 5:30 PM	**Registration**				
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
Genomics and Translational Bioinformatics Working Group Chairs: Ece Uzun, Wenyu Song		**Advanced Computational Statistics and Artificial Intelligence to Address Public Health Epidemics** Chairs: Naleef Fareed, Soledad Fernandez		**Microbiome Data Analysis: Advanced Methods and Practical Applications** Chairs: Qunfeng Dong, Xiang Gao	
8:30 AM - 8:50 AM	**Calibration of computational prediction tools for improved clinical variant classification and interpretation** Vikas Pejaver, Mount Sinai	8:30 AM - 8:50 AM	**Leveraging urinary drug test (UDT) results as a novel data source and proxy for drug use** Naleef Fareed/Soledad Fernandez, The Ohio State University	8:30 AM - 8:50 AM	**A Deep Learning Feature Importance Test Framework for Integrating Informative High-dimensional Biomarkers to Improve Disease Outcome Prediction** Baiming Zou, University of North Carolina at Chapel Hill
8:50 AM - 9:10 AM	**Opioid Prescriptions and Associated Patient Response: An Integrated Genetic**	8:50 AM - 9:10 AM	**Predicting opioid overdose mortality using UDT data with a Bayesian approach** John Myers, The Ohio	8:50 AM - 9:10 AM	**Enhancing Microbiome-Trait Prediction through Phylogeny-Aware Modeling and Data Augmentation**

	Analysis Using Clinical Biobank Wenyu Song, Brigham and Women's Hospital		State University		Yang Lu, University of Waterloo
9:10 AM - 9:30 AM	**Leveraging Deep Learning to Infer Cellular Dynamics** Shengyu Li, Houston Methodist Research Institute	9:10 AM - 9:30 AM	**Implementing a Spatial-Temporal Graph Neural Network (ST-GNN) framework, a novel, multi-modal data approach for predicting opioid overdose death rates** Xianhui Chen, The Ohio State University	9:10 AM - 9:30 AM	**Leveraging new genomic LLMs for studying under-annotated microbial genes** Siyuan Ma, Vanderbilt University Medical Center
9:30 AM - 9:50 AM	**Clinical and Genomic Investigation of Immune-Related Adverse Events** Qianqian Song, University of Florida	9:30 AM - 9:50 AM	**Addressing opioid-misuse abundance estimation challenges using Capture–Recapture methods** Fode Tounkara, The Ohio State University	9:30 AM - 9:50 AM	**Bayesian spatial statistical models for quantifying relationships among cell types in image data** Jacqueline R. Starr, Brigham and Women's Hospital, Harvard Medical School.
9:50 AM - 10:10 AM	**Machine Learning-Based Integration of Transcriptomic and Epigenetic Data for Cancer Biomarker Discovery** Alper Uzun, Brown University	9:50 AM - 10:10 AM	**Evaluating the public health decision support landscape for opioid outcomes** Brandon Slover/Neena Thomas, The Ohio State University	9:50 AM - 10:10 AM	**Multimedia: An R package for multimodal mediation analysis of microbiome data** Kris Sankaran, University of Wisconsin–Madison
10:10 AM -10:30 AM		***Coffee/Tea Break***			

10:30 AM - 10:50 AM	**Subtyping Metabolic Dysfunction-Associated Steatotic Liver Disease using Electronic Health Record-Linked Genomic Cohorts Reveals Diverse Etiologies and Progression** Shulan Tian, Mayo Clinic	10:30 AM - 10:50 AM	**Paper 46: PCORsearch: A Scalable, User-Centric Platform for Self-Service Cohort Discovery and Feasibility Analysis of PCORnet Data** Jacob Herman, The Ohio State University	10:30 AM - 10:50 AM	**VirusPredictor: Software to Predict Virus-related Sequences in Human Data** Dawei Li, Texas Tech University Health Sciences Center
10:50 AM - 11:10 AM	**Predicting Cancer Recurrence Using Deep Learning Based Models** Ece Uzun, Brown University	10:50 AM – 11:10 AM	**Paper 52: Towards AI Co-Scientists for Scientific Discovery in Precision Medicine** Hao Li, Washington University in St. Louis	10:50 AM - 11:10 AM	**Integrated Transcriptomics Analysis on Human Respiratory Viral Inoculation and Vaccine Challenge Studies** Fei Zou, School of Medicine, UNC
11:10 AM - 11:30 PM	**Genetic Impact of Alternative Transcription Initiation Reveals a Novel Molecular Phenotype for Human Diseases** Lei Li, Shenzhen Bay Laboratory	11:10 AM - 11:30 AM	**Paper 3: Tokenvizz: GraphRAG-Inspired Tokenization Tool for Genomic Data Discovery and Visualization** Zhenxiang Gao, Case Western Reserve University	11:10 AM - 11:30 AM	**AI-Powered Discovery of Novel Antimicrobial Peptides in *Trichomonas vaginalis*** Xiang Gao, Stritch School of Medicine, Loyola University Chicago
11:30 PM - 1:30 PM		**Lunch Break / Poster Session I**			

1:30 PM - 1:40 PM	**Opening Remarks (Room 320)**				
1:40 PM - 2:20 PM	**Keynote Lecture** (Room 320) **Veera Baladandayuthapani, PhD** **University of Michigan**				
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
Advancements in AI and Large Language Models for Biomedical Research Chairs: Jing Su, Gangqing Hu		**Big data for Better Studying Disease Systems** Chair: Xiuwei Zhang		**Advanced omics platforms and tools** Chairs: Kaixong Ye, Hongbo Liu	
2:30 PM - 2:50 PM	**Preliminary Evaluation of ChatGPT Model Iterations in Emergency Department Diagnostics** Gangqing Hu, West Virginia University	2:30 PM - 2:50 PM	**Eminent Scholar Presentation** Yu-Ping Wang, Tulane University	2:30 PM - 2:50 PM	**CCLLM: Cellular Community Large Language Model to identify motifs of cell organization in spatial transcriptomics** Juexin Wang, Indiana University
2:50 PM - 3:10 PM	**Thinking, Fast and Slow: DualReasoning Enhances Clinical Knowledge Extraction from Large Language Models** Jing Su, Indiana University School of Medicine	2:50 PM - 3:10 PM	**ShinyEvents: harmonizing longitudinal data for real world survival estimation.** Timothy Shaw, Moffitt Cancer Center	2:50 PM - 3:10 PM	**A universal gene representation of atlas single cell data** Hao Chen, University of Illinois Chicago

3:10 PM - 3:30 PM	**mcDETECT: Decoding the Dark Transcriptomes in 3D with Subcellular-Resolution Spatial Transcriptomics** Jian Hu, Emory University	3:10 PM - 3:30 PM	**Harnessing Big Data to Advance Understanding of Novel Therapeutic Strategies** Yuan Liu, Indiana University	3:10 PM - 3:30 PM	**Decoding Kidney Disease at Single-Cell Resolution: A Cross-Platform Spatial Transcriptomics Study** Haojia Wu, Washington University in Saint Louis
3:30 PM - 3:50 PM		***Coffee/Tea Break***			
3:50 PM - 4:10 PM	**A Visual-Omics Foundation Model for Integrating Histopathology Images and Transcriptomics** Weiqing Chen, Houston Methodist Research Institute	3:50 PM - 4:10 PM	**Spatially Resolved Transcriptomics and Proteomics to Interrogate Biological Mechanisms Underlying Cancer Disparities** Nina Steele, University of Cincinnati	3:50 PM - 4:10 PM	**DNA Methylation Predictors of Inflammatory Cytokine Changes in Breast Cancer Survivors Undergoing Chemotherapy** Hongying Sun, University of Rochester
4:10 PM - 4:30 PM	**Large language models in cancer pharmacogenomics: from drug-gene association to response prediction** Yuchiao Chiu, University of Pittsburgh	4:10 PM - 4:30 PM	**Studying single cells through multi-omics and multi-condition scRNA-seq** Xiuwei Zhang, Georgia Institute of Technology	4:10 PM - 4:30 PM	**Age-Related Patterns of DNA Methylation Changes** Gang Peng, Indiana University
4:30 PM - 4:50 PM	**STHD: probabilistic cell typing of single spots in whole transcriptome spatial data with high definition** Yi Zhang, Duke University	4:30 PM - 4:50 PM	**High-resolution reconstruction of single-cell specific spatial genome architectures in 3D space reveals context-specific mechanisms of long-range gene regulation**	4:30 PM - 4:50 PM	**Uncovering Hidden Biological and Technical Links from Large-scale DNA Methylome Data** Wanding Zhou, Children's Hospital of Philadelphia

			Jianrong Wang, Michigan State University		
4:50 PM - 5:10 PM	**Predicting Protein-Protein Interactions with Structure-based ML/DL Modeling** Haiqing Zhao, University of Texas Medical Branch	4:50 PM - 5:10 PM	**TBA** Jingwen Yan, Indiana University	4:50 PM - 5:10 PM	**Boosting Analysis Pipeline Efficiency in Bioinformatics Through Snakemake** Yaping Feng, Admera Health
5:10 PM - 5:30 PM	**A Benchmarking Framework for Foundation Models in Drug Response Prediction** Qianqian Song, University of Florida	5:10 PM - 5:30 PM	**Integrated Multi-Omics Study in Early Onset of Type 1 Diabetes.** Wenting Wu, Indiana University	5:10 PM - 5:30 PM	**A BLAST from the past: revisiting BLAST's E-value** Yang Lu, University of Waterloo

Monday, August 4th, 2025

8:30 AM - 6:00 PM	**Registration**				
8:30 AM - 9:10AM	**Keynote Speaker** (Room 320) **Jiang Bian, PhD** **Indiana University**				
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
Advances in target discovery and computational drug design Chairs: Pengyue Zhang, Yijie Wang		**Future Scientists in AI** Chair: Chi Zhang		**Bioinformatics Meet Biosignals: Opportunities and Challenges** Chairs: Haoqi Sun, Chen Huang	
9:20 AM - 9:40 AM	**Eminent Scholar Presentation**	9:20 AM - 9:40 AM		9:20 AM - 9:40 AM	**Bioinformatics Meets Biosignals:**

	Kaifu Chen Harvard Medical School				**Opportunities and Challenges** Haoqi Sun, Harvard Medical School
9:40 AM - 10:00 AM	**Dynamic Digital twins for early diagnosis and treatment** Mikael Benson, Karolinska Institutet	9:40 AM - 10:00 AM		9:40 AM - 10:00 AM	**Leveraging Clinical Biobanks and Genetics to Understand Sleep Apnea and Related Comorbidities** Brian Cade, Harvard Medical School
10:00 AM - 10:20 AM	**Drug repurposing for substance use disorders by genome-wide association studies and real-world data analyses** Dongbing Lai Indiana University	10:00 AM - 10:20 AM		10:00 AM - 10:20 AM	**Sleep Architecture Biomarkers of Psychiatric Disease** Shaun Purcell, Harvard Medical School
10:20 AM -10:40 AM		***Coffee/Tea Break***			
10:40 AM - 11:00 AM	**An Informatics Bridge to Improve the Design and Efficiency of Phase I Clinical Trials for Anticancer Drug Combinations** Lei Wang The Ohio State University	10:40 AM - 11:00 AM		10:40 AM - 11:00 AM	**Reimagining Sleep Medicine using AI-based Physiology-guided Digital Twins** Ankit Parekh, Icahn School of Medicine at Mount Sinai
11:00 AM- 11:20 AM	**Building an explainable graph neural network by sparse learning for the drug-protein binding prediction**	11:00 AM- 11:20 AM		11:00 AM- 11:20 AM	**Multi-omics in Neurodegenerative Diseases**

	Yijie Wang Indiana University				Bruno Benitez, Harvard Medical School
11:20 AM - 11:40 AM	**Combining genetics and real-world patient data fuel ancestry-specific target and drug discovery in Alzheimer's disease** Yuan Hou Cleveland Clinic	11:20 AM - 11:40 AM		11:20 AM - 11:40 AM	**Y-chromosome loss in cancer: single-cell insights into origins and consequences** Jun Xia, Texas A&M University
11:40 AM - 12:00 AM	**Identifying repurposable treatments in patient subpopulations** Pengyue Zhang, Indiana University	11:40 AM - 12:00 AM		11:40 AM - 12:00 AM	**Computational Techniques for Deciphering Cancer Genomics and the Tumor Microenvironment at Single-Cell Resolution** Jinzhuang Dou, The University of Alabama at Birmingham
12:00 AM - 12:20 PM	**So You Think You've Found a Target? Computational Simulation Methods for Hit Identification** Michael Robo, Indiana Biosciences Research Institute	12:00 AM - 12:20 PM		12:00 AM - 12:20 PM	**Distinct Signatures of Tumor-Associated Macrophages in Shaping Immune Microenvironment and Patient Prognosis** Chongming Jiang, Terasaki Institute for Biomedical Innovation
12:20 PM - 1:30 PM		**Lunch Break**			
1:30 PM - 2:10 PM		**Keynote Speaker** (Room 320) **Qing Nie, PhD** **University of California, Irvine**			

CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
AI and Applications for Better Understanding Disease Mechanisms Chair: Xubo Song		**Advances in Bioinformatics** Chairs: Yu-Chiao Chiu, Juexin Wang		**Integrative genomics and epigenomics to link GWAS variants to function** Chairs: Hongbo Liu, Kaixiong Ye	
2:20 PM - 2:40 PM	**Reprogramming Protein Language Models for Protein Function Annotation and Engineering** Yunan Luo, Georgia Institute of Technology	2:20 PM - 2:40 PM	**Eminent Scholar Presentation** Yufeng Shen, Columbia University	2:20 PM - 2:40 PM	**Precision Nephrology: The Role of Genetics in Kidney Health** Atlas Khan, Columbia University
2:40 PM - 3:00 PM	**MARVEL: Microenvironment Annotation by Supervised Graph Contrastive Learning** Yuying Xie, Michigan State University	2:40 PM - 2:50 PM	**Paper 20: Benchmarking Cellular Deconvolution Algorithms to Predict Cell Proportions: A Literature Review** Ayesha Malik, University of Central Florida	2:40 PM - 3:00 PM	**Unraveling the Molecular Heterogeneity of Severe Acute Malnutrition: Multi-omic Insights** Yixing Han, National Institutes of Health (NIH)
		2:50 PM - 3:00 PM	**Paper 30: Landscape of gene essentiality in cancer cell death pathways** Shangjia Li, The Ohio State University		
3:00 PM -		3:00 PM -	**Technology Session**	3:00 PM -	

3:20 PM	**Leveraging AI for Characterizing Pediatric Cancer** Shibiao Wan, University of Nebraska Medical Center	3:20 PM	**Boosting Pipeline Efficiency in Bioinformatics Through Snakemake** Shunian Xiang, Admera Health	3:20 PM	**Leveraging chromatin accessibility data to understand complex traits** Siming Zhao, Dartmouth College
3:20 PM -3:40 PM		***Coffee/Tea Break***			
3:40 PM - 4:00 PM	**Deep Learning models for image enhancement, translation, and harmonization** Xubo Song, Oregon Health & Science University	3:40 PM - 4:00 PM	**Spatial Transcriptomics at Scale with Stereo-seq: Big Data for Impactful Science** Yongfu Wang, Complete Genomics	3:40 PM - 4:00 PM	**Integrative genomics and epigenomics reveal functions of non-coding variants** Hongbo Liu, University of Rochester
4:00 PM - 4:20 PM	**Advancing AI for Individualized Diagnosis and Prognosis: From Prenatal Heart Defects to Prostate Cancer Survival** Jieqiong Wang, University of Nebraska Medical Center	4:00 PM - 4:20 PM	**Access the full richness of biological complexity with single cell and spatial multiomics from 10x Genomics** Nicole Jaymalin, 10x Genomics	4:00 PM - 4:20 PM	**Mechanistic annotation of GWAS loci for circulating fatty acids by single-cell omics and CRISPR screens** Huifang Xu, University of Georgia
4:20 PM - 4:40 PM	**TBA** Jordan Krull, Ohio State University	4:20 PM - 4:40 PM	**Directed Evolution of Molecular Enzymes Empowers NGS Library Preparation** Robin Song, Yeasen Biotechnology Co., Ltd.	4:20 PM - 4:40 PM	**Linking Rare Non-Coding Variants Associated with Human Longevity to Cellular Senescence via Integrative Functional Genomic Approaches** Jiping Yang, Columbia University Medical Center

4:40 PM - 5:00 PM	**Evaluate, standardize, and optimization bioinformatics software documentation using AI-agents** Shaopeng Gu, Ohio State University	4:40 PM - 5:00 PM	**Uncover Cellular Heterogeneity with Advanced Single Cell Multi-Omics Approaches** Julie Laliberte, Singleron Biotechnologies Inc., USA	4:40 PM - 5:00 PM	**Identification of replicative aging and inflammatory aging signatures via whole-genome CRISPRi screens and GWAS meta-analysis** Xueqiu Lin, Fred Hutchinson Cancer Center
5:00 PM - 6:00 PM	**Poster Session II (Atrium)**				

Tuesday, August 5th, 2025

8:30 AM - 6:30 PM	**Registration**				
8:30 AM - 9:10 AM	**Keynote Speaker** (Room 320) **Julie A. Johnson, PharmD** **The Ohio State University**				
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
Data science solutions for spatial transcriptomics Chair: Johnson Travis		**Computational Omics for Precision Medicine and Drug Discovery** Chairs: Bin Chen, Qian Li		**Integrative Bioinformatics for Translational and Precision Medicine** Chairs: Yuan Liu, Shilin Zhao	
9:20 AM - 9:40 AM	**Eminent Scholar Presentation** Xiang Zhou, Yale University	9:20 AM - 9:40 AM	**Protein Language Model ESM3 Enables Superior Prediction of Complex Variant Effects Across ClinVar and DMS Benchmarks** Xiaoming Liu, University of South	9:20 AM - 9:40 AM	**Paper 26: A novel immune-related risk stratification model to predict prognosis, immunotherapy and chemotherapy response for Neuroblastoma** Xiaohui Zhan

			Florida		
9:40 AM - 10:00 AM	**SpatialGE: An Interactive Web Platform for Accessible and Reproducible Spatial Transcriptomics Analysis** Xiaoqing Yu, Moffit Cancer Center	9:40 AM - 10:00 AM	**Massive labeled transcriptomics as a resource of transcriptome representation learning and drug discovery** Bin Chen, Michigan State University	9:40 AM - 10:00 AM	**Paper 1: The Impact of HLA Diversity on Immune Cell Composition, Tumor Mutation Burden, and Cancer Survival** Shilin Zhao, Vanderbilt University
10:00 AM - 10:20 AM	**Spatial Resolved Gene Regulatory Networks Analysis** Zhana Duren, Indiana University School of Medicine	10:00 AM - 10:20 AM	**Generative AI for Human Genetics and Functional Genomics** Xinghua (Mindy) Shi, Temple University	10:00 AM - 10:20 AM	**Paper 11: Horizontal gene transfer networks reveal resistance of plasmid-mediated communication in antibiotic exposure** Lijia Che, City Univesity of Hong Kong
10:20 AM - 10:40 AM		***Coffee/Tea Break***			
10:40 AM - 11:00 AM	**Identifying Key Regulators of Amyloid Beta Clearance from Single Cell Spatial Transcriptomics using Generalized Linear Mixed Effect Models** Debolina Chatterjee, Indiana University School of Medicine	10:40 AM - 11:00 AM	**Distinct Mutational Profiles in Primary Sclerosing Cholangitis-Associated Cholangiocarcinoma Compared to de novo Cholangiocarcinoma** Shulan Tian, Mayo Clinic	10:40 AM - 11:00 AM	**Paper 12: Boolean Network Modeling-Guided Identification of FDA-Approved Drug Combinations for Targeted Treatment Strategies in Head and Neck Cancer** Pranabesh Bhattacharjee, Texas A&M University
11:00 AM - 11:20 AM	**Leveraging Spatial Transcriptomics of**	11:00 AM - 11:20 AM	**High-resolution multi-omic**	11:00 AM - 11:20 AM	**Paper 27: Comparison of**

	Brain Tissue in Neurological Diseases Oscar Harari, The Ohio State University		**dissociation of brain tumors with multimodal autoencoder** Qian Li, Ph.D., St. Jude Children's Hospital		**Nanopore Sequencing, MethylationEPIC Array, and EM-Seq for DNA Methylation Detection** Steven Brooks, Indiana University
11:20 AM - 11:40 AM	**A Statistical Framework to Improve the Design of Spatial Transcriptomics Experiments** Dongjun Chung, The Ohio State University	11:20 AM - 11:40 AM	**CoMPaSS: A Computational Pipeline for Cross-Platform Concordance Assessment and Navigating Study Design in Microbiome Research** Xi Qiao, University of Utah	11:20 AM - 11:40 AM	**Paper 51: A Hierarchical Adaptive Diffusion Model for Flexible Protein-Protein Docking** Rujie Yin, Texas A&M University
11:40 AM - 12:00 AM	**Integrative Modeling of Gene Expression and Histology via Cross-Modal Alignment and Multi-Scale Graph Inference** Chao Chen, Stony Brook University	11:40 AM - 12:00 AM	**SEHI-PPI: An End-to-End Sampling-Enhanced Human-Influenza Protein-Protein Interaction Prediction Framework with Double-View Learning** Rui Yin, University of Florida	11:40 AM - 12:00 AM	**Paper 25: Knowledge-driven annotation for gene interaction enrichment Analysis** Lingxi Chen, City University of Hong Kong
12:00 AM - 12:20 AM	**Utilizing deep transfer learning to identify high risk subpopulations of cells in single cell spatial omics data**	12:00 AM - 12:20 AM	**Cyclin D1 induces epigenetic and transcriptional alterations in Multiple Myeloma with t(11;14)(q13;q32)**	12:00 AM - 12:20 AM	**Paper 22: "Frustratingly easy" domain adaptation for cross-species transcription factor binding prediction**

	Johnson Travis, Indiana University School of Medicine		Huihuang Yan, Mayo Clinic		Mark Maher Ebeid, University of Pittsburgh
12:20 PM - 1:30 PM		**Lunch Break**			
CONCURRENT SESSIONS					
Room: **320**		Room: **301**		Room: **350**	
AI and Machine Learning in Translational Genomics Chairs: Huihuang Yan, Yixing Han		**Data-Driven Insights into Disease Modeling** Chairs: Shulan Tian, Joseph McElroy		**Flash Talks** Chair: Zhifu Sun	
1:30 PM - 1:50 PM	**Paper 53: Adaptive Chebyshev Graph Neural Network for Cancer Gene Prediction with Multi-Omics Integration** Sa Li, Oakland University	1:30 PM - 1:50 PM	**Paper 48: Compositional Bayesian Co-Clustering of DTI biomarkers with Clinical Measures for Enhanced Prediction of Parkinson Disease Severity** Chandrajit Bajaj	1:30 PM - 1:40 PM	**Paper 45: A Multimodal Vision Transformer using Fundus and OCT Images for Interpretable Classifications of Diabetic Retinopathy** Shivum Telang, North Allegheny High School
				1:40 PM - 1:50 PM	**Paper 54: In Silico Design of a Population-Specific mRNA Vaccine Targeting MUC1 for Colorectal Cancer: Focus on Iranian HLA Diversity** Zarrin Minuchehr, National Institute of Genetic Engineering and Biotechnology

1:50 PM-2:10 PM	**Paper 2: A Generative Imputation Method for Multimodal Alzheimer's Disease Diagnosis** Reihaneh Hassanzadeh, Georgia Institute of Technology	1:50 PM-2:10 PM	**Paper 16: Latent factor modeling reveals unexpected spatial heterogeneity in human Alzheimer's disease brain transcriptomes** Hu Chen, Baylor College of Medicine	1:50 PM-2:00 PM	**Paper 28: Abnormal ERV expression and its clinical relevance in colon cancer** Aditya Bhagwate, Mayo Clinic
				2:00 PM-2:10 PM	**Paper 15: From Bench to Insight: Rapid Pathogen Genomic Surveillance Workflow for SARS-CoV-2 and Emerging Pathogens** Venkat Sundar Gadepalli, The Ohio State University
2:10 PM-2:30 PM	**Paper 13: A user-friendly R Shiny app for Predicting Surface Protein Abundance from scRNA-seq Expression Using Deep Learning in blood cells** Yidong Chen, University of Texas Health San Antonio	2:10 PM-2:30 PM	**Paper 31: DuAL-Net: A Hybrid Framework for Alzheimer's Disease Prediction from Whole-Genome Sequencing via Local SNP Windows and Global Annotations** Eun Hye Lee, Indiana University School of Medicine	2:10 PM-2:20 PM	**Paper 35: LoRA-BERT: a Natural Language Processing Model for Robust and Accurate Prediction of long non-coding RNAs** Nicholas Jeon, Texas A&M University
				2:20 PM - 2:30 PM	
2:30 PM-2:50 PM	**Paper 44: HELP-TCR Harmonized Explainable Language Processing toolkit for T-Cell antigen Receptor repertoires.** Yulyana Kalesnik,	2:30 PM-2:50 PM	**Paper 14: Resolving Gene Heterogeneity in DEG Analysis: A Novel Pipeline for Precision Genomics** Jiasheng Wang,	2:30 PM-2:40 PM	**Paper 36: ICM-MD: Integrating TM-Specific Features and MD-Derived Structures for Accurate Prediction of Inter-Chain Contacts**

	University of Lodz		Baylor College of Medicine		**in Alpha-Helical Transmembrane Homodimers** Bander Almalki, University of Delaware
				2:40 PM-2:50 PM	**Paper 24: OmicsSankey: Crossing Reduction of Sankey Diagram on Omics Data** Bowen Tan, City University of Hong Kong
2:50 PM-3:10 PM	**Paper 23: Efficient and Valid Large Molecule Generation via Self-supervised Generative Models** Doyoung Kwak, Texas A&M University	2:50 PM-3:10 PM	**Paper 29: Multimodal Imaging and Cell-Free DNA Methylation Analysis for Noninvasive Lung Cancer Diagnosis** Ran Hu, University of California – Los Angeles	2:50 PM-3:00 PM	**Paper 40: Multi-omic analysis integrating global transcriptional and post-transcriptional profiles reveals predominant role of post-transcriptional control in three human cell lines** Alexander Krohannon, Indiana University
				3:00 PM-3:10 PM	Abstract 22: **TCR Convergence as a Proxy for Tumor-Specific Immunity in HSV1-Positive rGBM Patients Treated with CAN-3110** Ayse Selen Yilmaz, The Ohio State University
3:10 PM - 3:30 PM		***Coffee/Tea Break***			

3:30 PM-3:50 PM	**Paper 47: DG-scRNA: Deep learning with graphic cluster visualization to predict cell types of single cell RNAseq data** Birkan Gokbag	3:30 PM-3:50 PM	**Paper 39: Multidimensional Impact of Microbiota Absence on Thymic T Cell Development in Mice: A Study Based on Single-Cell and Spatial Transcriptomics** Yifei Sheng, University of Chinese Academy of Sciences	3:30 PM-3:40 PM	Abstract 54: **Vritra: A Streamlined Pipeline for Species-resolved Functional Profiling of Target Genes in Microbiome Data** Boyan Zhou, New York University
				3:40 PM-3:50 PM	Abstract 66: **Supervised and Unsupervised Classification with Feature Selection for Single-Cell RNAseq Based on an Artificial Immune System** Dawid Krawczyk, University of Lodz
3:50 PM-4:10 PM	**Paper 50: A Machine Learning–Enhanced Pipeline for Detecting Disruption of Transcription Termination (DoTT) in RNA-Seq Data** Michael Levin, Temple University	3:50 PM-4:10 PM	**Investigating shared geospatial patterns in drug overdose behavior** Joanne Kim The Ohio State University	3:50 PM-4:00 PM	Abstract 76: **VaxLLM: An end-to-end framework leveraging a fine-tuned Large Language Model for automated vaccine annotation and database integration** Xingxian Li, University of Michigan
				4:00 PM-4:10 PM	Abstract 80: **Multi-Trait Polygenic Risk Score of Hypertension and Diabetes is Associated with Alzheimer's Disease Risk across Multi-Ethnic Cohorts** Anisha Das, Columbia University

4:10 PM- 4:30 PM	**THANOS: An AI Pipeline for Engineering Antibodies** Arnav Solanki	4:10 PM- 4:20 PM	**Paper 19: AutoRADP: An Interpretable Deep Learning Framework to Predict Rapid Progression for Alzheimer's Disease and Related Dementias Using Electronic Health Records** Qiang Yang, University Florida	4:10 PM- 4:20 PM	
		4:20 PM- 4:30 PM	**Paper 33: Machine Learning-Based Mortality Prediction in Critically Ill Patients with Hypertension: Comparative Analysis, Fairness, and Interpretability** Shenghan Zhang	4:20 PM- 4:30 PM	
4:30 PM- 4:40 PM	**Paper 7: DisSubFormer: A Subgraph Transformer Model for Disease Subgraph Representation and Comorbidity Prediction** Ashwag Altayyar, University of Delaware	4:30 PM- 4:40 PM	**Paper 37: Telehealth Utilization and Patient Experiences: The Role of Social Determinants of Health Among Individuals with Hypertension and Diabetes** Haoxin Chen	4:30 PM- 4:40 PM	

4:40 PM-4:50 PM	**Paper 32: GRN-Integrated Heterogeneous Attentive Graph Autoencoder for Cell-Cell Interaction Reconstruction from Spatial Transcriptomics** Aiwei Yang, Beijing Normal-Hong Kong Baptist University	4:40 PM-4:50 PM	**Paper 43: MetaphorPrompt2-A Structure and Function Focused Approach for Extracting Causal Events from Biological Text** Parth Patel, University of Texas at San Antonio	4:40 PM-4:50 PM	
5:00 PM - 5:45 PM		Award Presentation (Room 320)			
5:45 PM - 6:00 PM		Closing Remarks (Room 320)			

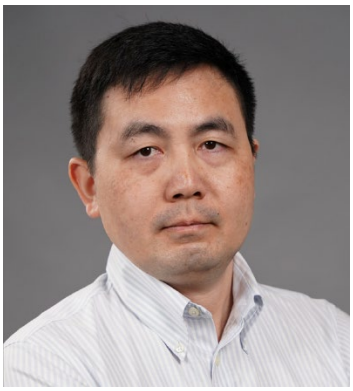
Keynote Speakers

Keynote Speaker
Veera Baladandayuthapani, Ph.D.
August 3rd
1:40 PM - 2:20 PM
Room: 320

Dr. Veera Baladandayuthapani is currently Jeremy M.G. Taylor Collegiate Professor and Chair in the Department of Biostatistics at University of Michigan (UM), where he also serves as the Associate Director of Associate Director of Quantitative Data Sciences and Director of the Cancer Data Science Shared Resource at UM Rogel Cancer Center. He obtained his Ph.D. in Statistics from Texas A&M University (in 2005), M. A. in Statistics from University of Rochester (in 2000) and BSc in Mathematics from Indian Institute of Technology, Kharagpur in 1998. He joined UM in Fall 2018 after spending 13 years in the Department of Biostatistics at University of Texas MD Anderson Cancer Center, Houston, Texas, where was a Professor and Institute Faculty Scholar and held adjunct appointments at Rice University, Texas A&M University and UT School of Public Health. His research explores the potential of Bayesian probabilistic models and machine learning methods to assist in medical and health sciences. These methods are motivated by large and complex datasets such as high-throughput genomics, epigenomics, transcriptomics and proteomics as well as high-resolution neuro- and cancer- imaging. A special focus is on developing integrative and spatial models combining different sources of data for biomarker discovery and clinical prediction to aid precision/translational medicine. His work has resulted in 160+ papers published in top statistical, biostatistical, bioinformatics, biomedical & oncology journals. He has also co-authored a book on Bayesian analysis of gene expression data. He has received several prestigious awards that include being selected as Myrto Lefkopoulou Distinguished Lectureship from Harvard School of Public Health; H. O. Hartley award from the Department of Statistics at Texas A&M University; Theodore. G. Ostrom from Washington State University; MD Anderson Faculty Scholar Award; Young Investigator Award from the International Indian Statistical Association (IISA) and Editor's Invited Paper for Biometrics, a top biostatistics journal and the flagship journal of the International Biometrics Society. He is a fellow of the American Association for Advancement in Science and the American Statistical Association and an elected member of the International Statistical Institute. He serves or has served on the Editorial board for major bio/-statistical journals such as Journal of American Statistical Association, Annals of Applied Statistics and Biometrics.

Title: Artificially Intelligent BioSpatial Modeling: Decoding Tumor Geography

Abstract: The tumor microenvironment (TME) is increasingly recognized as a critical frontier in cancer research, revealing how the spatial organization and dynamic interactions among diverse cell populations govern immune responses, tumor progression, and therapeutic outcomes. Recent advances in spatial profiling technologies—including spatial multiplex imaging, spatial transcriptomics, and digital pathology—have enabled unprecedented high-resolution characterization of these complex ecosystems. Yet these data introduce significant computational and statistical challenges: intricate spatial dependencies, substantial heterogeneity within and across samples, and non-conformable spaces that complicate integrative, population-level analyses. I will discuss my perspective on how the conflation of AI techniques and biologically-informed rigorous statistical modeling can address these challenges and unlock actionable biological insights. Specifically, I will discuss frameworks for modeling spatially varying genomic networks and transcriptional programs, approaches for quantifying intercellular interactions within the TME, and strategies for linking spatial features to patient-specific clinical outcomes. The utility and translational potential of these methods will be illustrated through multiple case studies spanning diverse cancer types.



Keynote Speaker
Jiang Bian, Ph.D.
August 4th
8:30 AM - 9:10 AM
Room: 320

Dr. Bian specializes in biomedical informatics and health data science, interdisciplinary fields focused on leveraging data, information, and knowledge to drive scientific discovery, problem-solving, and decision-making, all aimed at improving human health. Dr. Bian brings extensive experience in developing real-world data infrastructure, informatics tools, and systems, as well as applying advanced AI and data science methods to analyze and interpret multimodal clinical and biomedical data. Dr. Bian serves as Chief Data Scientist at Regenstrief, Chief Data Scientist at IU Health, and Associate Dean of Data Science among other leadership roles.

Title: Real-World Data to Real-World Evidence: Successes, Challenges, and Opportunities

Abstract: This presentation delves into the methods and tools enable the transformation of real-world data (RWD) into actionable real-world evidence (RWE). It emphasizes the central role of data science in addressing the inherent challenges of working with large-scale, messy, and heterogeneous data sources such as EHRs and claims data. Specific case studies—including target trial emulation for evaluating GLP-1 receptor agonists (GLP-1RAs) and outcomes in cancer risk and survivorship—demonstrate how advanced analytical frameworks and causal inference techniques can generate RWD complement randomized controlled trials. Also study design issues and target trial emulation.

Keynote Speaker
Qing Nie, Ph.D.
August 4th
1:30 PM - 2:10PM
Room: 320

Dr. Qing Nie is a University of California Presidential Chair and a Distinguished Professor of Mathematics and Developmental & Cell Biology at University of California, Irvine. Dr. Nie is the director of the *NSF-Simons Center for Multiscale Cell Fate Research* jointly funded by NSF and the Simons Foundation – one of the four national centers on mathematics of complex biological systems. In research, Dr. Nie uses systems biology and data-driven methods to study complex biological systems with focuses on single-cell analysis, multiscale modeling, cellular plasticity, stem cells, embryonic development, and their applications to diseases. Dr. Nie has published more than 250 research articles, including more than 50 papers in journals such as Nature, Science, Nature Methods, PNAS, Nature Machine Intelligence, Cancer Cells, Nature Communications. In training, Dr. Nie has supervised more than 60 postdoctoral fellows and PhD students, with many of them working in academic institutions. In 2025, Dr. Nie was ranked #1 by ScholarGPS based on citation metrics as *Highly Ranked Scholar* in two areas: a) Single-cell transcriptomics & b) Transcriptomics technologies for *Prior Five Years.* Dr. Nie has been recognized by various professional societies for his interdisciplinary research achievements. Dr. Nie is a fellow of the *American Association for the Advancement of Science (AAAS)*, *American Physical Society (APS)*, *Society for Industrial and Applied Mathematics (SIAM),* and *American Mathematical Society (AMS)*.

Title: Systems Learning of Single Cells

Abstract: Cells make fate decisions in response to dynamic environments, and multicellular structures emerge from multiscale interplays among cells and genes in space and time. While single-cell omics data provides an unprecedented opportunity to profile cellular heterogeneity, the technology requires fixing the cells, often leading to a loss of spatiotemporal and cell interaction information. How to reconstruct temporal dynamics from single or multiple snapshots of single-cell omics data? How to recover interactions among cells, for example, cell-cell communication from single-cell gene expression data? I will present a suite of our recently developed computational methods that learn the single-cell omics data as a spatiotemporal and interactive system. Those methods are built on a strong interplay among systems biology modeling, dynamical systems approaches, machine-learning methods, and optimal transport techniques. The tools are applied to various complex biological systems in development, regeneration, and diseases to show their discovery power. Finally, I will discuss the methodology challenges in systems learning of single-cell data.

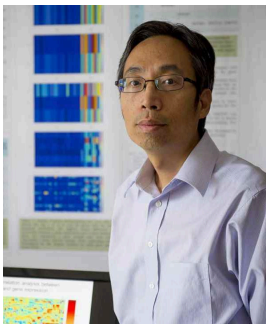
Keynote Speaker
Julie A. Johnson, Pharm.D.
August 5th
8:30 AM - 9:10 AM
Room: 320

Julie A. Johnson, Pharm.D. is the Dr. Samuel T and Lois Felts Mercer Professor of Medicine and Pharmacology at The Ohio State University's Colleges of Medicine and Pharmacy. She is the Director of OSU's Clinical and Translational Science Institute, Associate Dean for Research (Medicine) and Associate Vice President for Research at OSU. Dr. Johnson's research focuses on pharmacogenomics discovery and implementation and documenting outcomes of precision medicine approaches in clinical practice. She is an internationally recognized leader in clinical pharmacology, pharmacogenomics and genomic medicine, with over 340 peer reviewed original publications and over $55M in research funding as principal investigator, excluding the CTSA award. From 2015-2018 she was named a Clarivate Analytics Highly Cited Researcher, an accomplishment of about 1 in 1000 scientists globally. Dr. Johnson has received numerous awards and honors, including election to the National Academy of Medicine and election as fellow of the American Association for the Advancement of Science, and three other societies, along with top research awards from several multiple organizations. She was recently appointed to the National Academie's Forum on Drug Discovery, Development and Translational. She has received teaching awards from the University of Tennessee and the University of Florida and mentoring awards from the American Society for Clinical Pharmacology and Therapeutics and the American Heart Association.

Title: Using real world evidence to advance pharmacogenomics

Abstract: This presentation will cover the opportunities to advance discoveries and validation of genetic associations with efficacious or adverse reponses to drug therapy, including discussion of datasets in which this is possible. There will then be specific data presented from studies evaluating associations between CYP2D6 genotype and drug interactions in patients treated with opioid therapy and emergency department visits based on CYP2D6 phenotype status.

Eminent Scholar Talks



Eminent Scholar Talk

Yu-Ping Wang, Ph.D.
August 3rd
2:30 PM - 2:50 PM
Room: 301

Dr. Yu-Ping Wang received the BS degree in applied mathematics from Tianjin University, China, in 1990, and the MS degree in computational mathematics and the PhD degree in communications and electronic systems from Xi'an Jiaotong University, China, in 1993 and 1996, respectively. After his graduation, he had visiting positions at the Center for Wavelets, Approximation and Information Processing of the National University of Singapore and Washington University Medical School in St. Louis. From 2000 to 2003, he worked as a senior research engineer at Perceptive Scientific Instruments, Inc., and then Advanced Digital Imaging Research, LLC, Houston, Texas. In the fall of 2003, he returned to academia as an assistant professor of computer science and electrical engineering at the University of Missouri-Kansas City. He is currently a Professor of Biomedical Engineering, Computer Sciences, Neurosciences, and Biostatistics & Data Sciences at Tulane University. Dr. Wang's recent effort has been bridging the gap between biomedical imaging and genomics, where has over 250 journal publications. Dr. Wang is a fellow of AIMBE and won the 2022 Tulane Convergence Award for his effort in bridging gaps between science, engineering and biomedicine. He has served for numerous program committees and NSF and NIH review panels and is currently an associate editor for J. Neuroscience Methods, IEEE/ACM Trans. Computational Biology and Bioinformatics (TCBB) and IEEE Trans. Medical Imaging (TMI). More about his research can be found at his lab website: http://www.tulane.edu/~wyp/

Title: Integration of brain imaging and genomics with interpretable multimodal collaborative learning

Abstract: Recent years have witnessed the convergence of multiscale and multimodal brain imaging and omics techniques, showing great promise for systematic and precision medicine. In the meantime, they bring significant data analysis challenges when integrating and mining these large volumes of heterogeneous datasets. In this work, we first introduce a linear collaborative learning model to combine both regression and correlation analysis such as CCA. To further capture complex interactions both within and across modalities, we develop an interpretable multimodal deep learning-based integration model to perform heterogeneous data integration and result interpretation simultaneously. The proposed model can generate interpretable activation maps to quantify the contribution of imaging or omics features. Moreover, the estimated activation maps are class-specific, which can therefore facilitate the identification of biomarkers. Finally, we apply and validate the model in the study of brain development with integrative

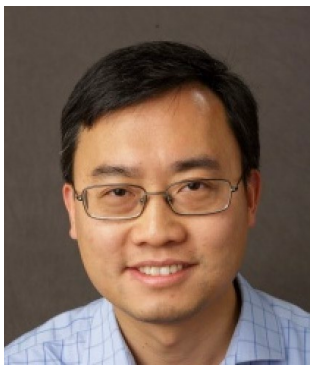
analysis of multi-modal brain imaging and genomics data. We demonstrate its successful application to both the classification of cognitive function groups and the discovery of underlying genetic mechanisms.

Eminent Scholar Talk
Kaifu Chen, Ph.D.
August 4th
9:20 AM -9:40 AM
Room: 320

Kaifu Chen, PhD, is currently an Associate Professor in the Pediatrics Department of Harvard Medical School and the Director of the Computational Biology Program in the Cardiology Department of Boston Children's Hospital. His research focused on understanding the expression regulation of cell identity genes by the binding of transcription factors to enhancers, histone modifications on nucleosomes, 3D genome folding, and cell-cell signaling in a tissue environment. His lab conducts bioinformatics analysis of multiomics data to understand these molecular mechanisms and develop AI models to uncover cell identity regulators based on these mechanisms, with a particular interest in applications to cardiovascular diseases and cancers.

Title: AI modeling of Cell Identity regulation in biological development and diseases

Abstract: Precise regulation of cell identity is fundamental in biological development and diseases, yet the underlying mechanisms remain incompletely understood. We present an integrated AI-driven framework that models cell identity regulation by combining large-scale single-cell transcriptomics, epigenomics, and regulatory network inference. Our analysis of millions of single cells across various human tissue types revealed that cell identity heterogeneity is shaped by both chromatin epigenetics and microenvironmental context. We developed a computational pipeline, MEBOCOST, to systematically characterize cell–cell communication and identified tissue-specific signaling patterns that modulate cell identity. To resolve the regulatory logic of cell identity, we introduced SCIG and CEFCIG—machine learning frameworks that uncover cell identity genes (CIGs) and their master regulators using genetic, epigenetic, and expression signatures. Applying this framework, we identified MECOM as a key regulator of endothelial lineage specification and demonstrated its role in enhancer looping, VEGF signaling, and angiogenesis. Together, our approach offers a unified model of cell identity regulation, advancing our understanding of tissue development, diseases, and regenerative interventions.



Eminent Scholar Talk
Yufeng Shen, Ph.D.
August 4th
2:20 PM - 2:40 PM
Room: 301

Yufeng Shen is an Associate Professor of Systems Biology and Biomedical Informatics at Columbia University. He received his B.Sc. in biochemistry from Peking University and his Ph.D. in computational biology from Baylor College of Medicine. At Baylor, he led the analysis of the first human genome sequenced by next-generation technologies. His research group is currently working on predicting effect of genetic variation using statistical and machine learning methods and to apply genomics and computational biology in genetic studies of human diseases. His group developed CANOES (Backenroth et al 2014) for calling copy number variants from exome sequencing data, gMVP (Zhang et al 2022), SHINE (Fan et al 2023), and MisFit (Zhao et al 2025) for predicting pathogenicity and fitness effect of protein variants. They discovered that epigenomic patterns in tissues under normal conditions are associated with risk genes of developmental disorders (Han et al 2018). In addition, his research led to the discovery of novel risk genes of congenital heart disease (Homsy et al 2015), congenital diaphragmatic hernia (Qi et al 2018, 2024), and autism (Zhou et al 2022).

Title: Representation and prediction of the impact of protein mutations

Abstract: Accurate prediction of genetic effect of missense variants is fundamentally important for disease gene discovery, clinical genetic diagnosis, personalized treatment, and protein engineering. Commonly used computational methods predict pathogenicity, which does not capture the quantitative impact on fitness in human. We developed a method, MisFit, to estimate selection coefficient of missense variants. MisFit jointly models the effect at a molecular level (D) and a population level (selection coefficient, S), assuming that in the same gene, missense variants with similar D would have similar S. We trained it by maximizing the probability of observed germline variant allele counts in 234,992 individuals of European ancestry. We show that S is informative in predicting allele frequency across ancestries and consistent with the fraction of de novo mutations observed in sites under strong selection. Further, S outperforms previous methods in prioritizing de novo missense variants in individuals with neurodevelopmental disorders. Finally, we show that predicted D and S are consistent with functional readout of deep mutational scan experiments of clinically important genes.

Eminent Scholar Talk
Xiang Zhou, Ph.D.
August 5th
9:20 AM -9:40 PM

Xiang Zhou is a Professor in the Department of Statistics and Data Science at Yale University. He earned a BS in Biology from Peking University in 2004, followed by an MS in Statistics (2009) and a PhD in Neurobiology (2010) from Duke University. He completed postdoctoral training in the Department of Human Genetics at the University of Chicago (2010–2013), where he later served as the Williams H. Kruskal Instructor in the Department of Statistics (2013–2014). Dr. Zhou joined the Department of Biostatistics at the University of Michigan as an Assistant Professor in 2014. He held the John G. Searle Assistant Professorship from 2018 to 2019 and was promoted to Associate Professor in 2019 and to full Professor in 2023. He served as Assistant Director of Precision Health (2022–2025) and, in 2025, became Assistant Director of Artificial Intelligence and Digital Health Innovation (AI&DHI). He joined Yale University in 2025. Dr. Zhou is a Fellow of the American Statistical Association and the recipient of the 2024 Mid-career Biosciences Faculty Achievement Recognition (MBioFAR) Award and the 2025 ICIBM Eminent Scholar Award. He is a standing member of the NIH MRAA Study Section and serves as an Associate Editor for PLOS Genetics, Journal of the American Statistical Association, and Annals of Applied Statistics. In 2024, he was Program Chair for the Section on Statistics in Genomics and Genetics of the American Statistical Association. His research centers on genomic data science, with a focus on developing advanced statistical and machine learning methods, including deep learning and AI tools, for the analysis of large-scale, high-dimensional genetic and genomic data. His work spans a range of application areas, including genome-wide association studies, single-cell sequencing, and spatial multi-omics.

Title: Statistical Methods for Single Cell Spatial Transcriptomics

Abstract: Spatial transcriptomics comprises a transformative set of genomic technologies that enable the measurement of gene expression with spatial localization information in tissue sections or cell cultures. In this talk, I will present several statistical methods recently developed by our group for analyzing spatial transcriptomics data. These include SPARK, a method for rigorous statistical detection of spatially variable genes (SVGs); SPARK-X, a fast and scalable approach for detecting SVGs in large-scale spatial transcriptomic studies; SpatialPCA, which enables spatially informed dimension reduction; CARD, a method for spatially guided cell type deconvolution; IRIS, a framework that integrates external single-cell data to support scalable spatial domain detection; and BASS, a hierarchical Bayesian model designed for multi-scale and multi-sample spatial transcriptomic analysis. Together, these methods provide a comprehensive toolkit for advancing the analysis and interpretation of complex spatial transcriptomic datasets.

Workshop – Genomics and Translational Bioinformatics

August 3rd
8:30 AM – 11:30 AM
Room: 320

Chairs: Ece Uzun, Wenyu Song

Title: Calibration of computational prediction tools for improved clinical variant classification and interpretation

Author list: Vikas Pejaver[1,2]
Detailed Affiliations:
[1]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

Abstract: The classification of genetic variants as being pathogenic or not is essential to the proper and timely diagnosis of genetic disorders. In 2015, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) provided formal guidance on how to weight and integrate different lines of evidence (e.g., population, functional, computational, among others) about a variant's pathogenicity or benignity towards its classification into one of five clinically relevant categories: *pathogenic*, *likely pathogenic*, *likely benign*, *benign* or of *uncertain significance*. As per these guidelines, evidence from computational and machine learning-based tools that use molecular and/or evolutionary information to predict the functional or phenotypic effects of variants, such as REVEL and AlphaMissense, was restricted to the weakest level of evidential strength (*Supporting*). However, the 2015 ACMG/AMP standards for the use of such predictors in variant classification and interpretation were based on developer-defined score thresholds, which are not always appropriate for the clinical context. Furthermore, these guidelines generally lacked quantitative support, predisposing them to being applied in non-standard ways that could lead to the misestimation of the evidential strength of *variant effect* predictors and inappropriate and/or inconsistent variant classification. To this end, we recently introduced a new calibration approach that standardizes scores from any variant effect predictor to 2015 ACMG/AMP evidence strength levels, from *Supporting* to *Very Strong*. Our approach estimates the local posterior probability of pathogenicity/benignity at a given prediction score to relate evidence strength as quantified by an existing Bayesian framework, with typical predictor performance measures such as precision and recall. Using carefully assembled independent data sets, we estimated score intervals corresponding to each level of evidential strength for pathogenicity and benignity for several different missense variant effect predictors and demonstrated that some predictors can reach up to *Moderate* and *Strong* evidence levels for a subset of variants. We then validated our proposed score intervals and estimated their impact on clinical variant classification using real-world data sets. Based on our findings, we recommended revisions of the ACMG/AMP criteria with respect to the use of variant effect predictors in the clinical context. Our work makes the use of such predictors in clinical variant classification and interpretation more rigorous and suggests a more prominent role for them in clinical genetic testing in the future.

Keywords: Variant effect predictor, calibration, variant classification, variant interpretation, ACMG/AMP guidelines, clinical genetic testing

Title: Opioid Prescriptions and Associated Patient Response: An Integrated Genetic Analysis Using Clinical Biobank

Author list: Wenyu Song[1, 10], Max Lam[2, 4], Ruize Liu[2, 3], Aurélien Simona[9], Scott G. Weiner[5, 10], Richard D. Urman[8], Kenneth J. Mukamal[6, 10], Adam Wright[7], David W. Bates[1, 10]

Detailed Affiliations:

1. Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. 2. Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. 3. Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston MA, USA. 4. Neurogenomics Laboratory @ IMH Research Division, Institute of Mental Health, Singapore. 5. Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. 6. Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. 7. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. 8. Department of Anesthesiology, The Ohio State University Wexner Medical Center, Columbus, OH, USA. 9. Division of Clinical Pharmacology and Toxicology, Geneva University Hospitals and Faculty of Medicine, Geneva, Switzerland. 10. Harvard Medical School, Boston, MA, USA

Abstract: Opioids are among the most powerful pain relievers available. Opioid drugs have been successfully used to treat both acute and chronic pain. While they can be effective for pain control especially acutely, opioids also have serious side effects and are prone to misuse and possible overdose. In 2023, there were 81,083 opioid related overdose deaths occurred in the United States. Genome-wide association studies (GWAS) have suggested that opioid related adverse events, including opioid use disorder (OUD), have strong genetic underpinnings. These genetic factors are located within genes that can affect efficacy, metabolism, and adverse effects of opioid drugs, which can in turn cause heterogeneous individual responses to drugs, including both pain levels and addiction. Electronic health records (EHRs) offer a largely untapped source of information to conduct genetic studies, which could facilitate the investigation of the genetic background of complex diseases and their comorbidities. EHR can be a particularly valuable data source for disorders like OUD that tend to be underrepresented in the cohort studies that comprise many genetic consortia.

We utilized patient-level clinical data from a large clinical biobank to develop opioid related phenotypes for genetic research. We first examined the genetic architecture of EHR-derived phenotypes of opioid use disorder (OUD) using GWAS and identified one novel significant OUD-associated locus on chromosome 4. Furthermore, we screened ~16 million rows of prescription records to develop codeine, one commonly prescribed opioid medicine, prescription-frequency phenotypes based on the number of recorded prescriptions for a given patient. Both low- and high-prescription counts were captured by developing 8 types of phenotypes with selected ranges of prescription numbers to reflect potentially different levels of opioid risk severity. We identified one significant locus associated with low-count codeine prescriptions (1, 2 or 3 prescriptions), while up to 7 loci were identified for higher counts, with a strong overlap across different thresholds. We identified 9 significant genomic loci with all-count phenotype. Further, using the polygenic risk approach, we identified a significant correlation between an externally derived polygenic risk score for opioid use disorder and numbers of codeine prescriptions. Our research provides a generalizable and clinical meaningful phenotyping pipeline for the genetic study of opioid-related risk traits.

Keywords: Electronic health record, Genome-wide association study, Opioid use disorder, Polygenic risk score, Opioid prescription phenotype

Title: Leveraging Deep Learning to Infer Cellular Dynamics

Author list: Shengyu Li[1,2,3,4], Pengzhi Zhang[1,2,3,4], Weiqing Chen[1,5], Lingqun Ye[1,2,3], Kristopher W. Brannan[2,3,4], Nhat-Tu Le[2,4], Jun-ichi Abe[6], John P. Cooke[2], Guangyu Wang[1,2,3,4,*]

Detailed Affiliations:

1. Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA. 2. Center for Cardiovascular Regeneration, Houston Methodist Research Institute, Houston, TX, USA. 3. Center for RNA Therapeutics, Houston Methodist Research Institute, Houston, TX, USA. 4. Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY, USA. 5. Department of Physiology, Biophysics & Systems Biology, Weill Cornell Graduate School of Medical Science, Weill Cornell Medicine, Cornell University, NY, USA. 6. The University of Texas MD Anderson Cancer Center, Department of Cardiology, Houston, TX, USA

Abstract: RNA velocity provides an approach for inferring cellular state transitions from single-cell RNA sequencing (scRNA-seq) data. Conventional RNA velocity models infer universal kinetics from all cells in an scRNA-seq experiment, resulting in unpredictable performance in experiments with multi-stage and/or multi-lineage transition of cell states where the assumption of the same kinetic rates for all cells no longer holds. Here we present cellDancer, a scalable deep neural network that locally infers velocity for each cell from its neighbors and then relays a series of local velocities to provide single-cell resolution inference of velocity kinetics. In the simulation benchmark, cellDancer shows robust performance in multiple kinetic regimes, high dropout ratio datasets and sparse datasets. We show that cellDancer overcomes the limitations of existing RNA velocity models in modeling erythroid maturation and hippocampus development. Moreover, cellDancer provides cell-specific predictions of transcription, splicing and degradation rates, which we identify as potential indicators of cell fate in the mouse pancreas.

Keywords: cell fate, RNA velocity, scRNA-seq, deep learning, cellDancer, relay velocity model

Title: Clinical and Genomic Investigation of Immune-Related Adverse Events

Author list: Yanfei Wang, PhD[1], Tyler A. Shugg, PharmD[2], Michael T. Eadon, MD[3], Jing Su, PhD[4], Thomas J George, MD[5], Jiang Bian, PhD[1], Steven M Smith[6], Yan Gong[7], Qianqian Song, PhD[1]

Detailed Affiliations:

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA**.** [2]Division of Clinical Pharmacology, Indiana University School of Medicine, Indianapolis, IN, USA**.** [3]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA**.** [4]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA**.** [5]Division of Hematology & Oncology, University of Florida & UF Health Cancer Center, Gainesville, FL, USA**.** [6]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, FL, USA**.** [7]Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Gainesville, FL, USA

Abstract: Immune checkpoint inhibitors (ICIs) have revolutionized cancer therapy, significantly improving survival outcomes across a range of malignancies. However, their use is frequently complicated by immune-related adverse events (irAEs), including acute kidney injury (AKI) and cardiovascular adverse events (CVAE), which can lead to treatment discontinuation and increased morbidity. To comprehensively characterize these risks, we leveraged large-scale clinical and genomic data from the OneFlorida+ Clinical Research Network and the All of Us Research Program. In a cohort of

ICI-treated patients from OneFlorida+, 56.2% developed irAEs within one year, with severe cases notably impacting overall survival. Younger patients, females, and those with specific comorbidities such as myocardial infarction and renal disease were at higher risk. Combination ICI regimens further increased irAE incidence, and cancer type also influenced risk profiles. In parallel, those ICI-treated patients revealed that 19.5% developed CVAEs, most commonly arrhythmias, myocardial infarction, and heart failure. Patients with pre-existing cardiometabolic conditions—including hypertension, diabetes, and hyperlipidemia—showed significantly elevated CVAE risk. Combination regimens, especially those involving CTLA-4 and PD-(L)1 inhibitors, were strongly associated with higher CVAE rates. Furthermore, genomic analysis of the All of Us cohort identified the rs16957301 variant in the PCCA gene as a novel genetic risk factor for ICI-AKI among Caucasian patients, with risk genotypes associated with earlier and higher incidence of AKI. Collectively, these findings highlight the critical importance of integrating clinical and genetic risk assessments to personalize ICI therapy, improve patient monitoring, and mitigate severe treatment-related toxicities. Tailored strategies addressing both renal and cardiovascular risks will be essential to ensure the safe and effective use of ICIs across diverse patient populations.

Keywords: Immune checkpoint inhibitors, Immune-related adverse events, risk factors

Title: Machine Learning-Based Integration of Transcriptomic and Epigenetic Data for Cancer Biomarker Discovery

Author list: Alper Uzun[1,2,3]

Detailed Affiliations:

[1]Department of Pathology and Laboratory Medicine, Warren Alpert Medical School of Brown University, Providence, RI 02903, USA; [2]Legorreta Cancer Center, Brown University, Providence, RI 02912, USA; [3]Brown Center for Clinical Cancer Informatics and Data Science (CCIDS), Brown University, Providence, RI 02912, USA

Abstract: ASCEND is a novel computational framework developed to integrate transcriptomic and DNA methylation data for accurate cancer prediction and biomarker discovery. Gene expression is a critical indicator of cellular function, and its dysregulation, often driven by epigenetic modifications like DNA methylation, plays a central role in cancer development. ASCEND bridges these two molecular layers to identify genes predictive of cancer and pinpoint methylation markers that may regulate their expression. Developed in Python using libraries such as scikit-learn, pandas, and numpy, ASCEND processes raw data from The Cancer Genome Atlas (TCGA), filters outliers and incomplete entries, and standardizes expression values across patients. Its workflow involves two main stages: the first predicts cancer presence using a Multilayer Perceptron Classifier trained on gene expression data to identify high-impact biomarker genes; the second uses a linear regression model to associate CpG methylation sites with those selected genes, revealing potential regulatory mechanisms. The tool automatically selects the top five genes based on importance scores but allows user customization. ASCEND was applied to datasets from breast, lung, and prostate cancers, comprising 370 samples from both healthy individuals and patients. In the case of breast adenocarcinoma, ASCEND achieved an 87% classification accuracy and identified WEE2P1, SUPT20HL1, TBC1D4, DGCR11, and TEX26 as the top candidate genes. Methylation analysis identified key CpG sites such as cg00396667, cg00493804, and cg00554640 from a pool of 27,577, many of which were mapped to these genes using ASCEND's feature importance scoring system. Literature review confirmed the relevance of four of the five identified genes to breast cancer, supporting

the model's reliability and biological relevance. ASCEND's results are visualized through a user-friendly interface, offering customizable parameters, graphical outputs, and modular adaptability to different cancer types. The tool is openly available on GitHub, supporting transparent, reproducible research. By linking gene expression and DNA methylation in a single platform, ASCEND offers researchers a scalable and insightful method for understanding the molecular basis of cancer, prioritizing diagnostic and therapeutic targets, and accelerating discoveries in precision oncology. Its design accommodates future expansion, enabling integration with additional omics layers and broader clinical applications, making ASCEND a valuable addition to the cancer bioinformatics toolkit.
Keywords: Cancer Biomarkers, Transcriptomics, DNA Methylation, Machine Learning, Multilayer Perceptron, Epigenetics

Title: Subtyping Metabolic Dysfunction-Associated Steatotic Liver Disease using Electronic Health Record-Linked Genomic Cohorts Reveals Diverse Etiologies and Progression
Author list: Tahmina Sultana Priya[1,7], Huihuang Yan[2,3], Kirk J. Wangensteen[4], Stephen Wu[2], Anthony C. Luehrs[5], Filippo Pinto e Vairo[3], Fan Leng[6], Andres J. Acosta[4], Robert, A. Vierkant[5], Alina M. Allen[4], Konstantinos N. Lazaridis[4], Eric W. Klee[2,3*], Danfeng (Daphne) Yao[1,7]*, Shulan Tian[2,3]*
Detailed Affiliations:
[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA; [2]Division of Computational Biology, Department of Quantitative Health Sciences, Rochester, MN, USA; [3]Center for Individualized Medicine and Department of Clinical Genomics, Mayo Clinic, Rochester, MN, USA; [4]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA; [5]Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA; [6]Data Analytics and Integration, Mayo Clinic, Rochester, MN, USA; [7]Sanghani Center for Artificial Intelligence and Data Analytics, Virginia Tech, Blacksburg, VA, USA;
Abstract: Metabolic dysfunction-associated steatotic liver disease (MASLD) is a heterogeneous condition with diverse etiologies and clinical presentations. Stratifying patients into homogenous subgroups, or subtypes, could potentially reveal molecular mechanisms driving disease risk and progression. Yet, a consensus of subtypes is lacking in MASLD, posing major challenges in developing tailored interventions. In this study, we developed a subtyping framework based on latent class analysis of significant MASLD-related clinical variables, followed by centroid-based assignment of new patients to established subtypes. We identified five subgroups with distinct genetic, clinical, and risk profiles, which were well recapitulated in an independent MASLD cohort. Polygenic risk score and genetic variant analysis revealed genetic contributions across all subgroups. In particular, two of the subgroups, male-predominant cardiorenal (C2) and female-predominant with obesity and mood disorders (C3), are associated with high prevalence of type 2 diabetes, obesity and sleep apnea. The latter also had relatively high usage of antidepressant medicine. On the other hand, the idiopathic subtype (C4) was characterized by the lowest incidence of metabolic comorbidities and ischemic heart disease. Nevertheless, this subgroup overall had the highest rate of liver transplant, which is likely driven, in part, by the combinatorial genetic effects of high-prevalent risk alleles in *TM6SF2* and *MBOAT7* together with low-prevalent protective allele in *HSD17B13*. Finally, the hepatic injury subtype C5 showed an increased risk

of developing advanced fibrosis and acute renal failure. Together, our study provides key insights into MASLD heterogeneity, highlighting the need for personalized therapies.

Keywords: Metabolic dysfunction-associated steatotic liver disease; Subtyping; Latent class analysis; Polygenic risk score; Precision medicine

Title: Predicting Cancer Recurrence Using Deep Learning Based Models

Author list: Jessica A. Patricoski-Chavez[1,2,3], Seema Nagpal[4], Ritambhara Singh[1,5], Jeremy L. Warner[2,6,7], Ece D. Gamsiz Uzun[1,2,3,6,7,8]

Detailed Affiliations:

[11]Center for Computational Molecular Biology, Brown University, Providence, RI; [2]Brown Center for Clinical Cancer Informatics and Data Science (CCIDS), Legorreta Cancer Center, Brown University, Providence, RI; [3]Department of Pathology and Laboratory Medicine, Brown University Health, Providence, RI; [4]Department of Neurology, Division of Neuro-oncology, Stanford University, Palo Alto, CA; [5]Department of Computer Science, Brown University, Providence, RI; [6]Departments of Medicine and Biostatistics, Brown University, Providence, RI; [7]Brown University Health Cancer Institute, Rhode Island Hospital, Providence, RI; [8]Department of Pathology and Laboratory Medicine, Warren Alpert Medical School of Brown University, Providence, RI

Abstract: Cancer is one of the leading causes of morbidity and mortality worldwide, with millions of new cases diagnosed each year. According to World Health Organization (WHO), nearly 10 million people worldwide lost their lives to cancer in 2020. Cancer recurrence remains a significant challenge, as it can occur months or even years after the initial treatment, underscoring the need for effective monitoring and predictive strategies. Understanding a patient's likelihood of cancer recurrence is crucial for optimizing treatment selection and timing, which can improve overall outcomes, and enhance quality of life. Deep learning (DL) models have shown increasing promise in various medical applications, including predicting disease recurrence. Gliomas represent approximately 25.5% of all primary brain and central nervous system (CNS) tumors and 80.8% of malignant brain and CNS tumors. Approximately 62% of patients experience recurrence within five years, and 17%–32% progress from low to high-grade glioma. Patients with low-grade gliomas (LGGs) have 5-year survival rates of up to 80%, while patients with higher-grade gliomas (HGGs) often experience rates below 5%. To explore the capability of DL models for predicting recurrence, we developed gLioma recUrreNce Attention-based classifieR (LUNAR), to predict early vs. late glioma recurrence using clinical, mutation, and mRNA-expression data from patients with primary grade II-IV gliomas from The Cancer Genome Atlas (TCGA). As an external validation set, we used the Glioma Longitudinal Analysis Consortium (GLASS). LUNAR outperformed all traditional ML models achieving area under the receiver operating characteristic curve (AUROC) of 82.84% and 82.54% on the TCGA and GLASS datasets, respectively.

Keywords: Cancer, deep learning, genomics, glioma, recurrence

Title: Genetic Impact of Alternative Transcription Initiation Reveals a Novel Molecular Phenotype for Human Diseases

Author list: Hui Chen[1], Xudong Zou[1], Wei Wang[2], Shuxin Chen[1], Yu Chen[2], Lei Li[1]

Detailed Affiliations:

[1]Shenzhen Bay Laboratory, Institute of Systems and Physical Biology. [2] Shenzhen Bay Laboratory, Institute of Cancer Research.
Abstract: Alternative transcriptional initiation (ATI) is a fundamental layer of gene regulatory mechanisms, characterized by multiple transcription start sites (TSSs) for a single gene, generating functionally distinct isoforms. ATI plays crucial regulatory roles in mediating tissue-specific gene expression and contributes significantly to the dysregulated transcriptomes observed in cancer. However, the genetic architecture underlying ATI variation and its mechanistic links to cancer susceptibility remain an important knowledge gap in the field. Here, we present the first comprehensive characterization of the genetic regulation of alternative transcription initiation (ATI) spanning 49 human normal tissues and 33 tumor tissues. We identified 9,075 5′UTR alternative transcription initiation trait loci (5′aQTLs), encompassing approximately 0.41 million common genetic variants associated with the usage of distal transcription start sites (TSSs) of 5,436 genes, 32.1% of which were overlooked by eQTLs. We found that 7.6% of disease variants are colocalized with 5′aQTL signals, and 74.0% of them were overlooked by eQTLs. By integrating our 5′aQTL with well-powered GWAS datasets through transcriptome-wide association studies (TWAS), we identified 156 cancer susceptibility genes, including established cancer markers such as *MAFF* and *MLLT10*, as well as novel candidates. Collectively, our study reveals ATI as a critical mechanism linking non-coding variants to cancer risk, providing new insight for cancer target discovery.
Keywords: alternative promoter; quantitative trait loci; transcriptome-wide association study; cancer susceptibility

Workshop – Advanced Computational Statistics and Artificial Intelligence to Address Public Health Epidemics

August 3rd
8:30 AM – 11:30 AM
Room: 301

Chairs: Naleef Fareed, Soledad Fernandez

Title: Leveraging urinary drug test (UDT) results as a novel data source and proxy for drug use

Author List: Naleef Fareed[1], Ping Zhang[1], Joanne Kim[1], Penn Whitley[2], Charles Mark[2], Brandon Slover[1], Steven Passik[2], Eric Dawson[2], John Myers[1], Xianhui Chen[1], Changchang Yin[1], Fode Tounkara[1], Neena Thomas[1], Bridget Freisthler[3], Tim Huerta[4], Soledad Fernandez[1], (Rebecca Jackson[5])
Detailed Affiliations:
[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC; [3] Department of Social Work, The Ohio State University; [4] Department of Family and Community Medicine, College of Medicine, The Ohio State University;; [5] Department of Internal Medicine, College of Medicine, The Ohio State University
Abstract: UDT data is example of a novel data source, like wastewater treatment data, to proxy fluctuating patterns of a public health phenomenon and support public health officials with decision making using signals to plan and predict crises. These signals could allow for better problem diagnosis, identification of

cold/hot clusters in a geographical area, and enable tailored responses to critically hit areas in a timely manner. Our first presentation provides context for the workshop by: 1) embedding the use of novel measures such as UDT to characterize differential human behavior and social processes; 2) discussing the scientific rationale for using novel measures such as UDT data within the context of the opioid crisis; and 3) describing the various analytical challenges of acquiring and processing multimodal and novel data sources for public health forecasting; and 4) lessons learned from our National Institute on Drug Abuse funded project to use UDT data, along with other multi-modal data sources, to develop a model for predicting opioid-related mortality outcomes. We will provide the audience with the logistics for the subsequent presentations and the learning objectives from each presentation. Drs. Fareed and Fernandez will field questions throughout the workshop as part of a Q&A session that will be held shortly after the presentations. There will also be a discussion of current challenges and limitations, along with strategic recommendations for the effective use of routinely collected data and emerging computational tools in forecasting public health trends and informing timely interventions.

Keywords: Prediction models; public health; epidemiology; artificial intelligence; community health; opioid crisis

Title: Predicting opioid overdose mortality using UDT data with a Bayesian approach

Authors: John Myers[1], Joanne Kim[1], Charles Marks[2], Penn Whitley[2], Brandon Slover[1], Naleef Fareed[1], Soledad Fernandez[1], Neena Thomas[1]

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC

Abstract: The opioid crisis represents an ongoing emergency in the United States, and Ohio has one of the highest overdose death rates in the country. Timely interventions are needed to combat this crisis; however, delays in overdose death reporting remain a significant hurdle in developing strategies to combat overdose deaths. Urine drug test (UDT) data provides near real-time weekly updates with valuable insights into drug use patterns in communities. We use UDT data as a proxy for drug use in communities in overdose death prediction and expect it to fill the gap in the lagged overdose death data.

We constructed a hierarchical Bayesian model implemented with Integrated Nested Laplace Approximation (INLA). The model allows spatiotemporal random effects to capture unobserved factors. Spatiotemporal effects were implemented with correlated random effects for space (Besag model) and time (first order random walk). UDT data, sociodemographic factors, and EMS events with naloxone distribution were used as parameters in the model.

Predictions were made at the quarterly level using the moving window approach to utilize the up-to-date drug overdose trend. We used a training-testing schema with a forgetting mechanism to train on eight quarters of data and predicted two quarters at a time, corresponding with the reporting lag of overdose death. We compared our model with baseline models to confirm that the predictive performance improved with the addition of UDT data and EMS naloxone events.

Keywords: Opioid overdose, Bayesian methods, Integrated Nested Laplace Approximation, Prediction, Urine Drug Test, Spatiotemporal

Title: Implementing a Spatial-Temporal Graph Neural Network (ST-GNN) framework, a novel, multi-modal data approach for predicting opioid overdose death rates

Author List: Zishan Gu[2], Xianhui Chen[2], John Myers[1], Joanne Kim[1], Changchang Yin[1], Naleef Fareed[1], Neena Thomas[1], Soledad Fernandez[1], Ping Zhang[2]

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Department of Computer Science and Engineering, College of Engineering, The Ohio State University

Abstract: The opioid crisis has severely impacted Ohio, with overdose death rates surpassing national averages and disproportionately affecting rural and Appalachian regions. Timely resource allocation and response are essential for preventing the further escalation of overdose deaths. Consequently, many studies focus on developing predictive models that enable timely interventions. Although these methods have demonstrated promising performance, there are three limitations: (1) they do not integrate the spatial relationships inherent in geographic information; (2) existing methods fail to integrate static information with dynamic data; and (3) current studies apply the same loss function to both large and small counties, leading to unfair modeling. To address these challenges, we propose the Spatial-Temporal Graph Neural Network (ST-GNN) framework, a novel approach for predicting opioid overdose death rates at county level. Our framework leverages the strengths of GNNs to model spatial relationships between counties, augmented by Long Short-Term Memory (LSTM) networks to capture temporal dynamics. In particular, this study uses Ohio's quarterly opioid overdose data from 2017 to 2023, enriched with dynamic features such as EMS naloxone administration events and static SDoH features, to train and evaluate the ST-GNN framework. By jointly training these components, the ST-GNN framework dynamically models the evolution of opioid overdose deaths across time and geography. Furthermore, to account for heterogeneity among counties with varying population sizes, we tailor prediction tasks accordingly with a joint training loss: for small counties where monthly or quarterly death counts are close to three, we formulate a binary classification task to predict whether the death count will exceed three. In contrast, for larger counties, we perform a standard regression task to predict the actual death counts. Compared with LSTM, DCRNN and GConvLSTM, our work not only advances the baselines in predictive modeling for opioid overdose deaths but also provides a scalable and flexible solution for addressing public health crises driven by complex spatial-temporal phenomena. Experiments conducted on data reported by the Ohio Department of Health demonstrate that our proposed method outperforms all baseline models for both large and small counties. For large counties, our method achieves an RMSE of 9.149 and an SMAPE of 0.242. For small counties, it achieves an ROC-AUC of 0.738 and an F1 score of 0.579.

Keywords: Opioid Overdose Prediction, Long Short-Term Memory (LSTM) networks, Graph Neural Network, Public Health

Title: Flexible Copula-Based Capture–Recapture Modeling of Opioid Misuse Using Urine Drug Testing Data: Evidence from Franklin County, Ohio (2016–2023)

Author List: Fode Tounkara[1], Naleef Fareed[1], Charles Marks[2], Penn Whitley[2], Neena Thomas[1], Soledad Fernandez[1]

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC;

Abstract: Understanding the hidden burden of opioid misuse remains a critical public health priority. Capture–recapture methods using urine drug testing (UDT) data offer a powerful framework for estimating the prevalence of people who misuse opioids (PWMO), especially when direct enumeration is not feasible. However, traditional models often fail to capture the complex heterogeneity and dependency structures among individuals.

We analyzed quarterly UDT data from Franklin County, Ohio, covering the years 2016 to 2023. Individuals were considered "captured" if they tested positive for any illicit opioid (e.g., heroin, fentanyl, oxycodone) during a given capture occasion. For each year, we derived binary capture histories and applied (1) a traditional generalized linear model (GLM) and (2) a suite of flexible copula-based zero-truncated binomial mixture models incorporating individual covariates (e.g., age). Copula families included Clayton, Gumbel, Joe, and Frank. For each model, we estimated the total number of PWMOs ($\hat{N}$), standard error (SE), 95% CI, and model fit (AIC). Subgroup analyses were conducted by sex.

Throughout the study period, the Frank copula model consistently outperformed or matched other models, especially when addressing asymmetric dependence structures. Estimates of $\hat{N}$ sometimes varied by over 20% between the Generalized Linear Model (GLM) and the best copula model, emphasizing the need to model latent correlations accurately. Subgroup analyses showed different trends in opioid misuse by gender, with females experiencing sharper increases in estimated prevalence post-2020, while male estimates remained more stable. Copula-based capture–recapture models provide a robust alternative for estimating hidden populations from complex surveillance data. By incorporating varied dependencies and individual covariates, our approach enhances opioid misuse prevalence estimates at the county level, informing public health resource allocation and overdose prevention strategies. This framework is also applicable to other substance use and surveillance areas.

Keywords: Opioid misuse, capture–recapture, copula model, urine drug testing, Franklin County, public health informatics

Title: Evaluating the public health decision support landscape for opioid outcomes

Author List: Naleef Fareed[1], Joanne Kim[1], Brandon Slover[1], Neena Thomas[1], Fernandez, S.

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University

Abstract: A data dashboard is an example of a digital tool that could effectively translate complex datasets into interactive visuals to support public health decision making. However, it can be very difficult to create a dashboard that provides both clear insight into a problem, and the ability to act on this insight. To do so, it is crucial to understand why the dashboard will be used, who will be using it, and what the desired impact of the dashboard will be. The overarching goal of our study is to develop a dashboard to communicate both statistical and machine learning prediction models to identify potential opioid overdose outbreaks across Ohio communities. By displaying these models on a dashboard along with simpler visualizations and filters, stakeholders in Ohio could gain some knowledge about the models and use the results to preemptively address opioid overdose outbreaks. As the dashboard was being built, questions emerged about what was necessary to create an effective tool for addressing real-world problems. What factors could be included that were the most impactful? What is the best way to arrange visualizations? How could the dashboard be created to hold the user's attention, and ensure it was not too difficult to use? It was decided that a more structured process was needed to develop an effective dashboard that would accomplish its desired goal. The research team systematically explored other existing state opioid dashboards to be leveraged as inspiration. First, a rubric was developed for grading eleven different components of each dashboard. Three

reviewers were then tasked with reviewing 42 of the state opioid dashboards. Cluster analysis was then performed to group the results based on the overall score, providing insight on which dashboards performed the best. Then, researchers met with the stakeholders who will be leveraging the dashboard to address the opioid overdose epidemic to prevent future outbreaks. These interviews were used to gauge what these individuals thought would be the most and least impactful to be included and ensure that all avenues of addressing the problem were considered. During our presentation, we will discuss our approach and how other researchers can adopt similar techniques to design and implement public health prediction tools using interactive tools such as dashboards.
Keywords: Dashboard, public health, statistical models, machine learning models, Ohio, opioids, overdose, prediction, review, cluster analysis, prevent, interviews, implement, interactive tools

Title: PCORsearch: A Scalable, User-Centric Platform for Self-Service Cohort Discovery and Feasibility Analysis of PCORnet Data
Author list: [1]Jacob Herman, [1]Maciej Pietrzak, [1]Neena Thomas, [1]Soledad Fernandez
Detailed Affiliations: [1]Department of Biomedical Informatics, College of Medicine, The Ohio State University
Abstract: Regulatory and institutional restrictions on electronic health record (EHR) access often cause delays, sometimes extending weeks, as researchers depend on independent data analysts for feasibility analysis. To address this bottleneck, we developed the Patient-Centered Outcomes Research Search Tool (PCORsearch), a web-based application enabling investigators to independently analyze deidentified EHR-derived data while maintaining regulatory compliance. PCORsearch allows interactive exploration of medical terminology codes and data definitions from the PCORnet Common Data Model, construction of custom queries, and retrieval of feasibility counts. Additionally, it supports cohort discovery, summary statistics, and visualization to characterize identified cohorts. By streamlining feasibility assessment and reducing reliance on data analysts, PCORsearch accelerates early-stage research planning and enables broader access to large-scale clinical data resources in a secure, compliant manner.
Keywords: Cohort Discovery; EHR Analytics; Research Feasibility; Feasibility Analysis; Research Automation; Deidentified Data; PCORnet; Patient-Centered Outcomes; EHR Tools; PCORI

Title: Towards AI Co-Scientists for Scientific Discovery in Precision Medicine
Author list: [1]Hao Li, [1]Di Huang, [1]Wenyu Li, [1]Heming Zhang, [1]Patricia Dickson, [1]J Philip Miller, [1]Carlos Cruchaga, [1]Michael Province, [1]Yixin Chen, [1]Philip Payne, [1]Fuhai Li
Detailed Affiliations:
[1]Washington University in St. Louis
Abstract: AI agents are emerging as transformative tools in precision medicine (AI4PM), tackling complex, poorly understood disease pathogenesis. We developed a multi-agent system coordinated by an Orchestrator agent that manages workflows, categorizes known facts, identifies gaps, and generates sequential task plans. This exploratory study demonstrates the system's potential to accelerate scientific discovery in AI4PM by structuring collaborative problem-solving in precision medicine research.
Keywords: Multi-agent; Medical agent; Ai-Medicine

Title: Tokenvizz: GraphRAG-Inspired Tokenization Tool for Genomic Data Discovery and Visualization
Author list: [1]Cerag Oguztuzun, [1]Zhenxiang Gao, [1]Jing Li, [1]Mehmet Koyuturk, [1]Rong Xu
Detailed Affiliations:
[1]Case Western Reserve University

Abstract: Interpreting complex genomic relationships and predicting functional interactions remain key challenges in biomedical research. Traditional sequence-based methods often lack interpretability which limits the exploration of genomic language model predictions. To address this gap, we introduce Tokenvizz, a GraphRAG-inspired tool that transforms genomic sequences into intuitive graph representations, where DNA tokens become nodes connected by edges weighted by attention scores derived from genomic language models. This novel approach translates genomic sequences into structured graph visualizations that reveal latent token relationships that are difficult to interpret through purely sequential methods. Tokenvizz provides an integrated pipeline that includes data preprocessing, graph construction from tokenized sequences, and an interactive web-based visualization interface. Users can dynamically adjust edge weight thresholds, perform position-based searches, and examine contextual sequence information interactively. This facilitates intuitive, multi-resolution analysis of genomic sequences and enhances the interpretability and exploratory capabilities of genomic language models.To validate Tokenvizz, we applied its graph representations to promoter-enhancer interaction prediction using a Graph Convolutional Network (GCN) on six datasets from the GUE+ benchmark. Tokenvizz consistently outperformed existing sequential deep learning models such as DNABERT2 and Nucleotide Transformer, demonstrating the utility of attention-derived graph structures for genomic prediction tasks. By effectively bridging attention-based genomic language modeling and interactive graph visualization, Tokenvizz offers researchers a visualization tool for exploratory genomic analyses. Future work will explore integrating external genomic annotation databases to further strengthen its interpretability and utility for genomics research. Tokenvizz, along with its user guide, is freely accessible on GitHub at: https://github.com/ceragoguztuzun/tokenvizz
Keywords: genomic language models; graph visualization; DNA sequence analysis; attention mechanism; regulatory elements; genomic interpretation

Workshop – Microbiome Data Analysis: Advanced Methods and Practical Applications

August 3rd
8:30 AM – 11:30 AM
Room: 350

Chairs: Qunfeng Dong, Xiang Gao

Title: A Deep Learning Feature Importance Test Framework for Integrating Informative High-dimensional Biomarkers to Improve Disease Outcome Prediction

Author List: Baiming Zou[1,2], James G. Xenakis[3], Meisheng Xiao[1], Apoena Ribeiro[4], Kimon Divaris[4], Di Wu[1,4], Fei Zou[1,5]

Detailed Affiliations:
[1]Department of Biostatistics, Giling School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA; [2]School of Nursing, University of North Carolina, Chapel Hill, NC, USA; [3]Department of Statistics, Harvard University, Cambridge, MA, USA; [4]School of Dentistry, University of North Carolina, Chapel Hill, NC, USA; [5]Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA.

Abstract: Many human diseases result from a complex interplay of behavioral, clinical, and molecular factors. Integrating low-dimensional behavioral and clinical features with high-dimensional molecular profiles can significantly improve disease outcome prediction and diagnosis. However, while some biomarkers are crucial, many lack informative value. To enhance prediction accuracy and understand disease mechanisms, it is essential to integrate relevant features and identify key biomarkers, separating meaningful data from noise and modeling complex associations. To address these challenges, we introduce the high-dimensional feature importance test (HdFIT) framework for machine learning models. HdFIT includes a feature screening step for dimension reduction and leverages machine learning to model complex associations between biomarkers and disease outcomes. It robustly evaluates each feature's impact. Extensive Monte Carlo experiments and a real microbiome study demonstrate HdFIT's efficacy, especially when integrated with advanced models like deep neural networks (DNN), termed HdFIT-DNN. Our framework shows significant improvements in identifying crucial features and enhancing prediction accuracy, even in high-dimensional settings.
Keywords: Complex association, Dimension reduction, Interpretable and scalable predictive modeling, Non-parametric feature selection, Stable deep neural network

Title: Enhancing Microbiome-Trait Prediction through Phylogeny-Aware Modeling and Data Augmentation
Author list: Yifan Jiang[1], Disen Liao[1],Matthew Aton[2], Qiyun Zhu[2], Yang Lu[1]
Detailed Affiliations:
[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada; [2]School of Life Sciences, Arizona State University, Tempe, AZ, USA.
Abstract: Understanding how microbial communities influence human traits is central to microbiome research and precision medicine. However, microbiome data present significant analytic challenges due to their high dimensionality, compositional constraints, and strong phylogenetic structure. In this talk, I will present two complementary methods that advance trait prediction from microbiome profiles by leveraging domain-specific priors: MIOSTONE, a taxonomy-aware neural network for interpretable prediction, and PhyloMix, a phylogeny-guided data augmentation technique. MIOSTONE improves interpretability and predictive accuracy by mimicking microbial taxonomy within the model architecture, enabling it to determine whether variations in microbial taxa at different taxonomic levels best explain the outcome. Complementarily, PhyloMix enhances learning by generating synthetic microbiome samples through subtree-level recombination guided by phylogenetic relationships. This strategy introduces meaningful diversity while respecting compositional constraints, boosting performance across multiple models and tasks, including supervised and contrastive learning. Together, these methods demonstrate the power of integrating biological structure into machine learning workflows for robust, interpretable, and effective microbiome-trait association studies.
Keywords: Microbiome-disease association; Biomarker detection; Taxonomy; Phylogeny; Data augmentation;

Title: Leveraging new genomic LLMs for studying under-annotated microbial genes
Author list: Siyuan Ma
Detailed Affiliations:
Department of Biostatistics, Vanderbilt University Medical Center

Abstract: Recent advancements in genomic large language models (LLMs) promise novel bioinformatics solutions for microbiome research. Microbial genomic sequences, like natural languages, form a *language of life*, enabling the adoption of LLMs to extract useful insights from complex microbial ecologies. In this talk, we will first review recent genomic LLMs, emphasizing their application towards metagenomics data. We will then present a particular application, namely, the study of unannotated microbial genes. We will demonstrate, with both bioinformatics evaluations and epidemiological findings based on public data, that novel genomic LLMs can be utilized to congregate otherwise unannotated microbial genes for more powerful downstream analysis and biologically interpretable findings.

Title: Bayesian spatial statistical models for quantifying relationships among cell types in image data
Author list: Jacqueline R. Starr[1]
Detailed Affiliations:
[1]Brigham and Women's Hospital, Harvard Medical School

Abstract: Our laboratory develops Bayesian spatial models to quantify and test relationships in biofilm or other FISH-based image data. I will describe these methods and how they can be used to investigate the role of bacteria (or other cells) in human health.

Title: Multimedia: An R package for multimodal mediation analysis of microbiome data
Author list: Hanying Jiang[1], Xinran Miao[1], Margaret W. Thairu[2], Mara Beebe[2], Dan W. Grupe[3], Richie J. Davidson[3,4,5], Jo Handelsman[6], Kris Sankaran[1]
Detailed Affiliations:
[1]Statistics Department, University of Wisconsin—Madison, Madison, Wisconsin, USA; [2]Wisconsin Institute for Discovery, University of Wisconsin—Madison, Madison, Wisconsin, USA; [3]Center for Healthy Minds, University of Wisconsin—Madison, Madison, Wisconsin, USA; [4]Psychology Department, University of Wisconsin—Madison, Madison, Wisconsin, USA; [5]Psychiatry Department, University of Wisconsin—Madison, Madison, Wisconsin, USA; [6]Plant Pathology Department, University of Wisconsin—Madison, Madison, Wisconsin, USA
Abstract: Mediation analysis has emerged as a versatile tool for answering mechanistic questions in microbiome research because it provides a statistical framework for attributing treatment effects to alternative causal pathways. Using a series of linked regressions, this analysis quantifies how complementary data relate to one another and respond to treatments. Despite these advances, existing software's rigid assumptions often result in users viewing mediation analysis as a black box. We designed the multimedia R package to make advanced mediation analysis techniques accessible, ensuring that statistical components are interpretable and adaptable. The package provides a uniform interface to direct and indirect effect estimation, synthetic null hypothesis testing, bootstrap confidence interval construction, and sensitivity analysis, enabling experimentation with various mediator and outcome models while maintaining a simple overall workflow. The software includes modules for regularized linear, compositional, random forest, hierarchical, and hurdle modeling, making it well-suited to microbiome data. Our case study revisits a study of the microbiome and metabolome of Inflammatory Bowel Disease patients, uncovering potential mechanistic interactions between the microbiome and disease-associated metabolites, not found in the original study. In addition to summarizing the package, we will explain the software design patterns that we drew inspiration from and how they could inform

reproducible multi-omics integration more generally. A gallery of examples and reference page can be found at https://go.wisc.edu/830110.
Keywords: Mediation analysis, data integration, multi-omics, microbiome, software design

Title: VirusPredictor: Software to Predict Virus-related Sequences in Human Data
Author list: Dawei Li[1]
Detailed Affiliations:
[1]Department of Immunology and Molecular Microbiology, Texas Tech University Health Sciences Center, Lubbock, TX 79430, USA
Abstract: Detecting disease-associated viruses without reference genomes, i.e., uncharacterized viruses, in human high-throughput sequencing data is challenging, as such sequences often evade alignment-based methods. Machine learning offers a promising alternative by classifying unmapped reads, potentially revealing novel viral elements. We developed VirusPredictor, a fast, open-source Python tool based on XGBoost and an in-house viral genome database. VirusPredictor uses a two-step classification approach: first, it categorizes sequences as infectious virus, endogenous retrovirus (ERV), or non-ERV human. Accuracy improves with sequence length, reaching 0.76 for short reads (150-350 bp, e.g., Illumina), 0.93 for mid-length reads (850-950 bp, e.g., Sanger), and 0.98 for long reads (2,000-5,000 bp). Sequences predicted as infectious viruses are then classified into one of six viral taxonomic subgroups, with accuracy increasing from 0.92 at 150-350 bp to >0.98 at >850 bp. These results suggest that assembling short reads into contigs (>1,000 bp) enhances prediction accuracy. VirusPredictor achieved high performance on real-world genomic and metagenomic datasets. To our knowledge, this is the first machine learning framework to incorporate both ERV classification and viral subgroup prediction, offering a practical solution for characterizing unmapped sequences potentially derived from uncharacterized viruses.
Keywords: XGBoost; Alignment-free prediction; Unmapped sequences; Uncharacterized virus; Endogenous retrovirus

Title: Integrated Transcriptomics Analysis on Human Respiratory Viral Inoculation and Vaccine Challenge Studies
Author list: Fei Zou
Detailed Affiliations:
Department of Genetics, School of Medicine, UNC
Abstract: Respiratory viral infections cause significant acute and chronic illness and substantial healthcare and economic burden. Human inoculation and vaccine challenge studies offer a unique opportunity to monitor immune cell responses with a controlled timeline. In this talk, I will first present HR-VILAGE-3K3M, the largest human respiratory viral immunization longitudinal gene expression repository database that we have built with publicly accessible transcriptomics data. I will then describe a set of integrated analyses that we perform on HR-VILAGE-3K3M to demonstrate its utility and to investigate cell mediated systemic and local immunity responses to viral inoculation and vaccine challenges.

Title: AI-Powered Discovery of Novel Antimicrobial Peptides in *Trichomonas vaginalis*
Author list : Qunfeng Dong[1], Xiang Gao[1]

Detailed Affiliations:
Department of Medicine, Stritch School of Medicine, Loyola University Chicago, 2160 S 1st Ave, Maywood, IL 60153
Abstract: Antimicrobial peptides (AMPs) are key components of innate immunity but are challenging to identify using traditional sequence-based methods due to their structural diversity. To address this, we fine-tuned ESM-2, a BERT-style protein language model, on a balanced dataset of AMP and non-AMP sequences. The fine-tuned model demonstrated strong performance in distinguishing AMPs and was applied to a large set of uncharacterized protein sequences from the NCBI RefSeq database. Among the candidates identified, many were from *Trichomonas vaginalis*, a protozoan pathogen known to disrupt vaginal microbial balance by suppressing *Lactobacillus* species. Additional analyses, including transcriptomic data and genomic context, suggest that a substantial portion of these candidates are expressed and associated with mobile genetic elements, pointing to possible roles in host interaction and adaptation. These findings highlight the potential of AI-driven approaches to uncover novel antimicrobial peptides and offer new perspectives on parasite biology and pathogenesis.

Workshop – Advancements in AI and Large Language Models for Biomedical Research

August 3rd
2:30 PM – 5:30 PM
Room: 320

Chairs: Jing Su, Gangqing Hu

Title: Preliminary Evaluation of ChatGPT Model Iterations in Emergency Department Diagnostics
Author list: Jinge Wang[1], Kenneth Shue[1], Li Liu[2,3], Gangqing Hu[1]
Detailed Affiliations:
[1]Department of Microbiology, Immunology & Cell Biology, West Virginia University, Morgantown, WV 26506, USA; [2]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA; [3]Biodesign Institute, Arizona State University, Tempe, AZ, 85281 USA.
Abstract: Large language model chatbots such as ChatGPT have shown the potential in assisting health professionals in emergency departments (EDs). However, the diagnostic accuracy of newer ChatGPT models remains unclear. This retrospective study evaluated the diagnostic performance of various ChatGPT models—including GPT-3.5, GPT-4, GPT-4o, and o1 series—in predicting diagnoses for ED patients (n=30) and examined the impact of explicitly invoking reasoning (thoughts). Earlier models, such as GPT-3.5, demonstrated high accuracy for top-three differential diagnoses (80.0% in accuracy) but underperformed in identifying leading diagnoses (47.8%) compared to newer models such as chatgpt-4o-latest (60%, $p < 0.01$) and o1-preview (60%, $p < 0.01$). Asking for thoughts to be provided significantly enhanced the performance on predicting leading diagnosis for 4o models such as 4o-2024-0513 (from 45.6% to 56.7%; $p = 0.03$) and 4o-mini-2024-07-18 (from 54.4% to 60.0%; $p = 0.04$) but had minimal impact on o1-mini and o1-preview. In challenging cases, such as pneumonia without fever, all models

generally failed to predict the correct diagnosis, indicating atypical presentations as a major limitation for ED application of current ChatGPT models.
Keywords: ChatGPT; large language models; emergency medicine; diagnosis; model iterations

Title: **Thinking, Fast and Slow: DualReasoning Enhances Clinical Knowledge Extraction from Large Language Models**
Author list: Haining Wang[1], Chenxi Xiong[1,2], Suthat Liangpunsakul[3], Wanzhu Tu[1], Jing Su[1]
Detailed Affiliations:
[1]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indiana, IN, USA; [2]Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA; [3]Department of Medicine, Indiana University School of Medicine, Indiana, IN, USA
Abstract: Large language models (LLMs), trained on vast repositories of medical knowledge, present transformative opportunities for clinical machine learning. However, their potential for extracting actionable insights from unstructured real-world data remains underexplored. We introduce DualReasoning, a novel framework that leverages LLMs to extract meaningful clinical information from medication records, enhancing predictive modeling and disease phenotyping. DualReasoning integrates deliberative, slow (Chain-of-Thought, CoT) and instinctive, fast (non-CoT) reasoning, reflecting human cognitive processes. We applied this approach to the All of Us cohort (N=247,652), which included 26,987 Type 2 Diabetes (T2D) cases, to extract diabetes-related knowledge from medication records. The extracted knowledge was used to enhance downstream predictive models for T2D phenotyping, including logistic regression, random forest, XGBoost, and multi-layer perceptron (MLP). These models incorporated demographic characteristics and Charlson comorbidities as predictors. DualReasoning's outperforms conventional feature engineering approaches (i.e., Polydrug risk scores [PdRS]) and LLM-based medication embeddings. By synergizing slow and fast reasoning, DualReasoning enhances clinical knowledge extraction through better use of medication information. This hybrid framework not only improves disease phenotyping but also shows promise for analyzing complex behavioral patterns, social determinants of health, and mental health conditions—areas where traditional approaches often fall short. Future work will explore its adaptability to broader clinical applications.
Keywords: large language model, drug informatics, knowledge extraction, knowledge presentation, phenotyping

Title: mcDETECT: Decoding the Dark Transcriptomes in 3D with Subcellular-Resolution Spatial Transcriptomics
Author list: Chenyang Yuan[1,2], Krupa Patel[1], Hongshun Shi[1,3,4], Hsiao-Lin V. Wang[1,5], Feng Wang[1], Ronghua Li[1,5], Yangping Li[1†], Victor G. Corces[1,5], Hailing Shi[1,4,5], Sulagna Das[1,6], Jindan Yu[1,3,4], Peng Jin[1,5], Bing Yao[1 *] & Jian Hu[1,2 *]
Detailed Affiliations:
1. Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA. 2. Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA. 3. Department of Urology, Emory University School of Medicine, Atlanta, GA 30322, USA. 4. Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA 30322, USA. 5. Emory Center for Neurodegenerative Diseases, Emory University School of Medicine, Atlanta,

GA 30322, USA. 6. Department of Cell Biology, Emory University School of Medicine, Atlanta, GA 30322, USA
Abstract: Spatial transcriptomics (ST) has shown great potential for unraveling the molecular mechanisms of neurodegenerative diseases. However, most existing analyses of ST data focus on bulk or single-cell resolution, overlooking subcellular compartments such as synapses, which are fundamental structures of the brain's neural network. Here we present mcDETECT, a novel framework that integrates machine learning algorithms and *in situ* ST (iST) with targeted gene panels to study synapses. mcDETECT identifies individual synapses based on the aggregation of synaptic mRNAs in three-dimensional (3D) space, allowing for the construction of single-synapse spatial transcriptome profiles. By benchmarking the synapse density measured by volume electron microscopy and genetic labeling, we demonstrate that mcDETECT can faithfully and accurately recover the spatial location of single synapses using iST data from multiple platforms, including Xenium, Xenium 5K, MERSCOPE, and CosMx. Based on the subsequent transcriptome profiling, we further stratify total synapses into various subtypes and explore their pathogenic dysregulation associated with Alzheimer's disease (AD) progression, which provides potential targets for synapse-specific therapies in AD progression.
Keywords： spatial transcriptomics, RNA granule, subcellular structure, Alzheimer’s disease, machine learning

Title: A Visual-Omics Foundation Model for Integrating Histopathology Images and Transcriptomics

Author list: Weiqing Chen[1,5,#], Pengzhi Zhang[1,2,3,4,#], Guangyu Wang[1,2,3,4]
Detailed Affiliations:
1. Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, 77030, USA. 2. Center for Cardiovascular Regeneration, Houston Methodist Research Institute, Houston, TX, 77030, USA. 3. Center for RNA Therapeutics, Houston Methodist Research Institute, Houston, TX, 77030, USA. 4. Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY, 10065, USA. 5. Department of Physiology, Biophysics & Systems Biology, Weill Cornell Graduate School of Medical Science, Cornell University, New York, NY, 10065, USA
Abstract: Artificial intelligence has revolutionized computational biology, particularly with the emergence of omics technologies such as single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST), which provide detailed genomic insights alongside tissue histology. However, existing computational models typically focus on either omics- or image-based analysis, lacking an integrated approach.
To bridge this gap, we developed OmiCLIP, a visual-omics foundation model that links hematoxylin and eosin (H&E) images with transcriptomic data using tissue patches from Visium datasets. For transcriptomic representation, we generated ‘sentences’ by concatenating the top-expressed gene symbols from each tissue patch. We compiled a dataset of 2.2 million paired tissue images and transcriptomic profiles across 32 organs to train OmiCLIP, enabling a robust integration of histology and transcriptomics.
Building upon OmiCLIP, we developed the Loki platform, which offers five core functionalities: (1) tissue alignment, (2) tissue annotation using bulk RNA-seq or marker genes, (3) cell type decomposition, (4) image–transcriptomics retrieval, and (5) ST gene expression prediction from H&E images. Compared with 22 state-of-the-art models across five simulated datasets, 19 public datasets, and four in-house experimental datasets, Loki consistently demonstrated superior accuracy and robustness across all tasks.

Keywords: large language model, contrastive learning, histology images, omics

Title: Large language models in cancer pharmacogenomics: from drug-gene association to response prediction

Author list: Yu-Chiao Chiu[1]

Detailed Affiliations:

[1]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA

Abstract: Large language models (LLMs) are emerging as powerful tools in pharmacogenomics, driving both qualitative and quantitative advances in cancer drug studies. This talk highlights two recent applications. First, we introduce a retrieval-augmented framework that qualitatively infers drug-gene-cancer associations by synthesizing evidence from PubMed literature. This efficient approach supports a pan-cancer interaction network and a web-based inference tool. Second, we present a quantitative strategy using an ensemble machine learning model that integrates molecular fingerprints with LLM-derived drug embeddings to predict responses in polyploid giant cancer cells (PGCCs), a chemoresistant breast cancer subpopulation. Together, these studies highlight the potential of LLMs to advance cancer pharmacogenomics through accelerated discovery and predictive modeling.

Keywords: Large language models; Pharmacogenomics; Drug-gene interactions; Cancer drug response

Title: STHD: probabilistic cell typing of single spots in whole transcriptome spatial data with high definition

Author list: Chuhanwen Sun[1]*, Yi Zhang[1,2,3,4,5]*#

Detailed Affiliations:

[1]Department of Neurosurgery, Duke University; [2]Department of Biostatistics and Bioinformatics, Duke University; [3]Department of Cell Biology, Duke University; [4]Brain Tumor Omics Program, The Preston Robert Tisch Brain Tumor Center, Duke University; [5]Duke Cancer Institute. *These authors contributed equally.

Abstract: Recent spatial transcriptomics (ST) technologies have enabled single- and sub-cellular resolution profiling of gene expression across the whole transcriptome. However, the transition to high-definition ST significantly increased data sparsity and dimensionality, posing computational challenges in identifying cell types, deciphering neighborhood structure, and detecting differential expression - all are crucial steps to study normal and disease ST samples. Here we present STHD, a novel machine learning method for probabilistic cell typing of single spots in whole-transcriptome, high-resolution ST data. Unlike the current binning-aggregation-deconvolution strategy, STHD directly models gene expression at single-spot level to infer cell type identities without cell segmentation or spot aggregation. STHD addresses sparsity by modeling count statistics, incorporating neighbor similarities, and leveraging reference single-cell RNA-seq data. We show in VisiumHD data that STHD accurately predicts cell type identities at single-spot level, which achieves precise segmentation of both global tissue architecture and local multicellular neighborhoods. The high-resolution labels facilitate various downstream analyses, including cell type-stratified bin aggregation, spatial compositional comparisons, and cell type-specific differential expression analyses. Moreover, STHD labels further reveal frontlines of inter-cell type interactions at immune hubs in cancer samples. STHD is scalable and generalizable across diverse samples, tissues, diseases, and different spatial technological platforms, facilitating genome-wide analyses in various spatial organization contexts. Overall, computational modeling of individual spots with STHD

facilitates discoveries in cellular interactions and molecular mechanisms in whole-genome spatial technologies with high resolution. STHD is available at https://github.com/yi-zhang/STHD/.
Keywords: Machine learning, spatial transcriptomics, high definition

Title: Predicting Protein-Protein Interactions with Structure-based ML/DL Modeling
Author list: Haiqing Zhao[1,2], Zhiyuan Song[1,2], Diana Murray[3], Barry Honig[3]
Detailed Affiliations:
[1]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA; [2]Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX, USA; [3]Department of Systems Biology, Columbia University Medical School, New York, NY, USA
Abstract body: Predicting whether two proteins physically interact has become a central challenge in computational biology. While recent deep learning (DL) approaches have shown promise in predicting protein–protein interactions (PPIs), they often lack the computational efficiency required to interrogate the vast number of possible interactions across the proteome. To address this, we developed **PrePPI-AF**, an algorithm that integrates structural information and additional sources of biological evidence to predict PPIs across most of the human proteome. PrePPI-AF leverages AlphaFold-predicted structures, which are parsed into individual domains to construct potential interaction models. In parallel, we introduced **ZEPPI** (Z-score Evaluation of Protein-Protein Interfaces), a framework that evaluates structural models of protein complexes using residue-level sequence co-evolution within interface regions. ZEPPI demonstrated superior performance over other deep learning–based approaches, particularly in evaluating models from the CASP-CAPRI benchmark experiments. We integrated the PrePPI and ZEPPI pipelines with a protein language model-based method to predict the *E. coli* PPI interactome. Clustering the resulting high-confidence predictions revealed functionally coherent subnetworks—even though our methods incorporated no explicit functional annotations. Together, these findings suggest that our proteome-wide prediction framework can serve as an efficient large-scale screening tool, which can be followed by more computationally intensive structural modeling for specific PPIs of interest.
Keywords: Protein-Protein Interaction, Protein Structure Prediction

Title: A Benchmarking Framework for Foundation Models in Drug Response Prediction
Author list: Qing Wang[1,2], Yining Pan[1], Minghao Zhou[1], Qianqian Song[1]
Detailed Affiliations:
[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32611, USA
Abstract: Understanding and overcoming drug resistance remains a critical challenge in improving cancer treatment outcomes. Advances in single-cell technologies have enabled high-resolution profiling of cellular heterogeneity, offering new insights into therapeutic response. At the same time, large foundation models are rapidly transforming computational biology, yet their applications in predicting drug response using single-cell data remain underexplored. Our work has introduced an integrated benchmarking framework designed to evaluate the performance of diverse foundation models for drug response prediction. This framework incorporates ten foundation models, including models trained on single-cell and general language data, across a curated data resource from a wide range of tissues, cancers, and treatment conditions. Our results demonstrate that model performance varies substantially depending on

the evaluation scenario and adaptation strategy. Certain models achieve high accuracy when fine-tuned on labeled data, while others maintain strong generalization in zero-shot settings without retraining. These findings not only provide practical guidance for selecting appropriate models for specific research or clinical contexts, but also highlight the diverse strengths of different modeling paradigms. By offering a flexible, extensible, and user-accessible framework, our study lays a critical foundation for advancing AI-driven drug discovery, supporting the broader goal of enhancing therapeutic decision-making and improving patient outcomes.
Keywords: Single-cell Profiling, Foundation Models, Drug Response Prediction, Low-Rank Adaptation, Zero-shot Learning, Computational Drug Discovery

Workshop – Big data for Better Studying Disease Systems

August 3rd
2:30 PM – 5:30 PM
Room: 301

Chairs:

Workshop – Advanced omics platforms and tools

August 3rd
2:30 PM – 5:30 PM
Room: 350

Chairs: Kaixiong Ye, Hongbo Liu

Title: CCLLM: Cellular Community Large Language Model to identify motifs of cell organization in spatial transcriptomics
Author list: Chunyang Chai[1], Yang Yu[2], Shuang Wang[3], Dong Xu[2], Huiyan Sun[1], Juexin Wang[4]
Detailed Affiliations:
[1]School of Artificial Intelligence, Jilin University, Changchun, 130012, China; [2]Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA; [3]Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47405, USA; [4]Department of Biomedical Engineering and Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, Indianapolis, IN 46202, USA
Abstract: The organization of diverse cellular types and states is recognized to be associated with tissue function. However, spatial principles and underlying mechanisms governing that organization remain largely unresolved across most physiological and pathological contexts. Detecting explicit conserved patterns of spatial cell organization as topological cell type combinations, known as Cellular Community motifs (CC motifs), suffer from high computational costs and limited detection accuracy. We introduce Cellular Community Large Language Model (CCLLM) to identify CC motifs leveraging fine-tuned large

language models (LLMs) modeling graphs constructed from spatial transcriptomics data. By converting spatial cellular distributions into structured textual prompts, CCLLM accurately identifies and counts subgraph patterns within cellular communities. We apply CCLLM on synthetic and real-world datasets to show its effectiveness and robustness in identifying disease-specific CC motifs with varying spatial resolutions, pathological conditions, and treatment responses. CCLLM harnesses the reasoning capabilities of LLMs to generate biologically meaningful interpretations of CC motif functions. This framework underscores the potential of graph-based LLMs modeling biological systems, offering insights into cellular communication dynamics and therapeutic target discovery.
Keywords: Spatial transcriptomics, cellular community, large language model, graph modeling, motifs

Title: A universal gene representation of atlas single cell data
Author list: Hao Chen[1,2], Nam D. Nguyen[1], Matthew Ruffalo[1], Ziv Bar-Joseph[1]
Detailed Affiliations:
[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA; [2]Department of Computer Science, University of Illinois Chicago, IL, USA.
Abstract: Recent efforts to generate atlas-scale single cell data provide opportunities for joint analysis across tissues and across modalities. Most of the existing methods for single cell atlas analysis use cells as the reference unit to combine datasets. However, such methods suffer from the limitations of inability to effectively integrate cross-modality data, hindering downstream gene-based analysis, and loss of genuine biological variations. Here we present a new data integration method, GIANT, which is for the first time designed for the atlas-scale analysis from the gene perspective. GIANT first converts datasets from different modalities into gene graphs, and then recursively embeds genes in the graphs into a latent space without additional alignment. Applying GIANT to the HuBMAP datasets creates a unified gene embedding space across multiple human tissues and data modalities, where gene representations reflect the functions of genes in their cells. Further evaluations demonstrate the usefulness of GIANT in discovering diverse gene functions, and underlying gene regulations in cells of different tissues.
Keywords: Single cell, Spatial transcriptomics, Representation learning, Multi-modal

Title: Decoding Kidney Disease at Single-Cell Resolution: A Cross-Platform Spatial Transcriptomics Study
Author list: Haojia Wu[1], Benjamin D. Humphreys[1,2]
Detailed Affiliations:
[1]Division of Nephrology, Department of Medicine, Washington University in St. Louis School of Medicine, St. Louis, MO, USA; [2]Department of Developmental Biology, Washington University in St. Louis School of Medicine, St. Louis, MO, USA
Abstract: Recent advances in spatial transcriptomics have transformed our understanding of how cells are organized in space and how their interactions affect organ function and dysfunction. This understanding is especially important for studying complex organs such as the kidney, where disease progression depends not only on cell-intrinsic changes but also on spatial relationships within the renal microenvironment. While multiple platforms are available for generating high-resolution spatial transcriptomic profiles, it remains unclear which technologies are best suited for studying the kidney and kidney diseases. In this study, we focus on commercially available platforms that provide single-cell resolution spatial transcriptomics data and compare the performance of four technologies, including full

transcriptome profiling platforms (Stereo-seq and VisiumHD) and targeted gene panel platforms (Xenium and MERFISH), in both healthy mouse kidneys and those affected by acute kidney injury. All platforms demonstrated strong ability to identify major kidney cell types at single-cell resolution. However, missed cell types were observed in targeted gene panel platforms (Xenium and MERFISH) when key marker genes were absent from the panel design. Stereo-seq and VisiumHD had comparable sensitivity in transcript and gene detection, although Stereo-seq uniquely detected long non-coding RNAs that were not captured in VisiumHD data. On the other hand, VisiumHD offers the advantage of integrating high-quality histological staining images aligned to the spatial transcriptomics data, which can provide enhanced context for downstream data interpretation. When comparing Xenium and MERFISH, Xenium consistently produced cleaner cell clustering and more robust results. While MERFISH detected a higher number of transcripts per cell, its signal distribution was noisier than Xenium. Accurate gene panel selection was found to be critical for both platforms in capturing diverse cell types and cellular states. We have determined the minimal number of genes required to reliably identify all kidney cell types using targeted platforms. Furthermore, we evaluated gene imputation and data integration tools to enhance analysis when only limited gene sets were available. Finally, by integrating data from Stereo-seq, Xenium, and MERFISH in the context of acute kidney injury, we identified disease-associated microenvironments and spatial gene expression shifts that were not captured in our previous single-cell RNA-seq analysis of the same disease model.
Keywords: Spatial transcriptomics, Kidney diseases, VisiumHD, StereoSeq, Xenium, MERFISH

Title: DNA Methylation Predictors of Inflammatory Cytokine Changes in Breast Cancer Survivors Undergoing Chemotherapy
Author list: Hongying Sun[1], Michelle C. Janelsins[1,2]
Detailed Affiliations:
[1]Department of Surgery, Division of Supportive Care in Cancer, University of Rochester Medical Center, Rochester, NY, USA; [2] Wilmot Cancer Institute, Rochester, NY, USA
Abstract: Background: Inflammation is a critical mechanism driving cancer-related cognitive dysfunction, fatigue, and other long-term adverse effects in breast cancer survivors. However, predictive biomarkers for post-treatment inflammatory burden remain limited. DNA methylation, an epigenetic marker that reflects environmental and biological exposures, may serve as an early predictor of systemic inflammation following chemotherapy. **Methods:** We analyzed data from 241 participants enrolled in a multi-center observational cohort: 192 breast cancer patients treated with chemotherapy and 49 age- and sex-matched controls. Whole blood DNA methylation was profiled using Illumina HumanMethylation450 and EPIC BeadChip arrays at three timepoints: baseline (T1), post-chemotherapy (T2), and a change score representing within-person differences (T2_1). A panel of inflammatory cytokines—including IL-6, IL-10, IL-4, IL-8, TNF-α, soluble TNF receptors I and II (sTNFRI, sTNFRII)—was measured from plasma and log-transformed. Epigenome-wide association studies (EWAS) were conducted using linear regression models adjusted for treatment group and array type. Each CpG site was modeled as a predictor of cytokine concentration at follow-up timepoints. False discovery rate (FDR < 0.05) was used to identify statistically significant associations. **Results:** The most robust epigenome-wide associations were observed for soluble TNF receptor II (sTNFRII) at outcome timepoint T2_1 (post-treatment change). Baseline (T1) methylation predicted T2_1 sTNFRII levels with 2,475 significant CpGs (FDR < 0.05), while follow-up (T2) and change (T2_1) methylation yielded 36 and 843 significant CpGs, respectively. Notably, many of the top differentially methylated CpGs mapped to immune- and inflammation-related

genes, including *TNFRSF1B*, *IL1RAP*, *SOCS3*, and *NFKBIA*. In contrast, other cytokines such as IL-6, IL-10, and IL-8 showed fewer significant CpG associations (generally fewer than 25 CpGs per model). No substantial associations were observed in the control group, suggesting chemotherapy-specific epigenetic responses. **Conclusions:** Our findings demonstrate that DNA methylation at baseline is strongly associated with downstream inflammatory activation, particularly for sTNFRII, a marker of TNF-α signaling and immune aging. These results suggest that methylation profiles prior to chemotherapy may serve as predictive biomarkers for inflammation-related late effects in cancer survivors. Future work will focus on validating these epigenetic signals in independent cohorts and integrating multi-omics data to understand causal pathways of immune dysregulation.
Keywords: DNA methylation, cancer-related cognitive impairment, breast cancer, chemotherapy, inflammation

Title: Age-Related Patterns of DNA Methylation Changes
Author list: Kevin Chen[1], Wenshu Wang[2], Hari Naga Sai Kiran Suryadevara[3], Gang Peng[4]
Detailed Affiliations:
[1]DeBakey High School, Houston, TX, USA; [2]Herricks High School, New Hyde Park, NY, USA; [3]Bioinformatics Core, Pennington Biomedical Research Center, Louisiana State University, Baton Rouge, LA, USA; [4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA
Abstract: DNA methylation is a dynamic process that adds methyl groups to DNA molecules, which plays an important role in gene expression regulation, genome stability, and aging. Epigenetic clocks developed over the past decade have used DNA methylation values at specific CpG sites combined with factors such as environmental influences and lifestyle factors to predict biological age with remarkable accuracy. However, these epigenetic clocks, built using conventional penalized regression models, have several limitations: (1) they assume consistent linear changes in methylation with age, even though methylation patterns may vary at different ages; (2) the selected CpGs in these clocks often lack specific biological significance; and (3) there is minimal overlap among the selected CpGs across different clocks, suggesting that only a small subset of a larger group of age-correlated CpGs is used to build an effective clock. Using data from 4,899 samples across 23 GEO datasets, we first analyzed the methylation patterns of 1,868 CpGs from 9 widely used epigenetic clocks and observed the expected consistently linear patterns in a subcluster. Next, we applied our own CpG selection method, which does not assume a lifelong linear correlation between DNA methylation and age. we divided the lifespan into overlapping age windows— [0, 20], [5, 25], [10, 30] …, [55, 75], [60, 80]— and identified CpGs that showed strong correlations with age within each specific window. Most CpGs were selected in early and later life windows, while few were selected in the middle life windows, indicating fast DNA methylation changes during child development and aging. Among the 12,903 unique CpGs selected within any of the windows, we performed clustering and uncovered from main patterns: (1) increase before 20 years old, (2) decrease before 20 years old, (3) increase before 20 and then decrease after 65, and (4) decrease before 20 and then increase after 65. Comparing this to the clock CpGs, both sets contained CpGs that decreased during adolescence. When examining gender differences, we noted that female samples had notably slowed methylation changes in old age windows compared to male samples, possibly due to hormonal changes during menopause. These findings suggest that age-related methylation changes are more complex than previously thought, with implications for refining epigenetic clocks and redefining the way researchers

study aging. Future research should explore the biological mechanisms behind these non-linear and sex-specific patterns.
Keywords: DNA methylation, Biological Age, Sex-specific

Title: Uncovering Hidden Biological and Technical Links from Large-scale DNA Methylome Data
Author list: David Goldberg[1], Sol Moe Lee[1], Cameron Cloud[1], Wanding Zhou[1,2]
Detailed Affiliations:
[5]Children's Hospital of Philadelphia, Philadelphia, PA, USA,
[2]Department of Pathology and Lab Medicine, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, USA
Abstract: Epigenome-wide association studies (EWAS) are transforming our understanding of the interplay between epigenetics and complex human traits and phenotypes. We performed scalable and quantitative screening of trait-associated DNA cytosine modifications in larger, more inclusive, and stratified human populations. We profiled the ternary-code DNA methylations—dissecting 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and unmodified cytosine—yielding multiple biological insights. We revealed a previously unappreciated role of 5hmC in trait associations and epigenetic clocks. We demonstrated that 5hmCs complement 5-methylcytosines (5mCs) in defining tissues and cells' epigenetic identities. In-depth analyses highlighted the cell type context of EWAS and GWAS hits. Using multiple platforms, we conducted a comprehensive human 5hmC aging EWAS, discovering tissue-invariant and tissue-specific aging dynamics, including distinct tissue-specific rates of mitotic hyper- and hypomethylation. These findings chart a landscape of the complex interplay of the two forms of cytosine modifications in diverse human tissues and their roles in health and disease.
Keywords: Epigenetics; DNA methylation; EWAS; Hydroxymethylation

Title: Boosting Analysis Pipeline Efficiency in Bioinformatics Through Snakemake
Author list: Hua ke[1], Jingling Hou[1], Shunian Xiang[1], Nihir Patel[1], Yaoqi Li[1], Haixin Shu[1], Si Chen[1], Yaping Feng[1]
Detailed Affiliations:
[1]Department of Bioinformatics, Admera Health, New Jersey, NJ, USA
Abstract: Given the complexity and diversity of Bioinformatics (BI) analyses, automation has become increasingly challenging, especially for biotechnology corporations that must process a high volume of daily projects involving multiple species and custom client requirements. Automating BI workflows through the integration of scripts, pipelines, and AI-powered systems enables a more standardized and scalable approach, significantly improving the ability to manage large datasets, accelerating analysis speed, enhancing reproducibility, minimizing user input and human error, and ultimately allowing researchers to focus on scientific discovery in biology and medicine.
Snakemake has proven to be an outstanding workflow management system for building scalable and maintainable pipelines. We are leveraging Snakemake to develop a modular bioinformatics analysis package that supports a wide range of next-generation sequencing (NGS) data types-including bulk RNA-seq, small RNA-seq, microRNA-seq, single-cell RNA-seq, and spatial transcriptomics. Each pipeline is structured into independent, reusable modules that can be flexibly combined via a simple command-line interface, accommodating customized analytical requests.
For example, a typical bulk RNA-seq workflow may involve modules such as sequencing quality control, adapter trimming, genome alignment, quantification (gene/isoform-level), differential expression analysis,

gene ontology enrichment, KEGG pathway mapping, and gene set enrichment analysis. Our system can automatically assemble a Snakemake workflow by selecting any combination of these modules. This modular design not only supports flexible pipeline configuration but also allows for targeted optimization, seamless integration with version-controlled repositories like GitHub, and collaborative development across teams.
Keywords: Snakemake, Workflow management, Reproducibility, Automation

Title: A BLAST from the past: revisiting BLAST's E-value
Author list: Yang Lu[1], William Stafford Noble[2], Uri Keich[3]
Detailed Affiliations:
[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON,; [2]Department of Genome Sciences and Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA; [3]School of Mathematics and Statistics, University of Sydney, Camperdown, NSW, Australia.
Abstract: The Basic Local Alignment Search Tool, BLAST, is an indispensable tool for proteomic and genomic research. BLAST's E-values have provided scientists with a meaningful statistical evaluation of reported sequence similarity searches over the past 30 years. Here we critically reevaluate these E-values, showing that they can be at times significantly conservative while at others too liberal. We offer an alternative approach based on generating a small sample from the null distribution of random optimal alignments and testing whether the observed alignment score is consistent with it. In contrast with BLAST, our significance analysis seems valid through extensive simulated and real data experiments. One advantage of our approach is that it works with any reasonable choice of substitution matrix and gap penalties, avoiding BLAST's limited options of matrices and penalties. In addition, we can formulate the problem using a canonical family-wise error rate control setup, thereby dispensing with E-values, which can at times be difficult to interpret.
Keywords: BLAST; E-value; Sequence Comparison

Workshop – Advances in target discovery and computational drug design

August 4th
9:20 AM – 12:20 PM
Room: 320

Chairs: Pengyue Zhang, Yijie Wang

Title: Dynamic Digital twins for early diagnosis and treatment

Author list: Mikael Benson

Detailed Affiliations:
Medical Digital Twin Research Group, Karolinska Institutet, Sweden

Abstract: Digital twins is a concept from engineering, which has been applied complex systems

such as airplanes or even cities. The key idea is to computationally model those systems, in order to develop and test them more quickly and economically than in real life. STDC applies this concept to personalize medicine, by constructing 1) network models of all molecular, phenotypic and environmental factors relevant to disease mechanisms in individual patients (digital twins); and 2) computationally treat those twins with thousands of drugs in order to identify the best one or ones to 3) treat the patient.

Keywords: Digital twins

Title: Drug repurposing for substance use disorders by genome-wide association studies and real-world data analyses

Author list: Dongbing Lai

Detailed Affiliations:
Indiana University

Abstract: Substance use disorders (SUDs, including alcohol, cannabis, opioids, nicotine, etc.) represent significant public health challenges and the prevalence is rising rapidly. Many individuals with SUDs use more than one substance and they often experience increased risks of overdose, mental health problems, and chronic diseases, and interactions of multiple substances complicate diagnosis and treatment. Drugs treating SUDs are available; however, their efficacy is limited, and relapse rates of SUDs remain high. Studies have shown that genetic factors contribute to ~50% variations of SUDs and there exist genes responsible for multiple SUDs (SUD-shared genes). Repurposing drugs targeting SUD-shared genes provides an efficient and effective way to develop novel drugs to treat SUDs, especially for those using multiple substances. In this study, we conducted the largest genome-wide association studies of SUDs to date to identify SUD-shared genes using samples from European, African, and Latinx ancestries (N=1,683,739). We innovatively considered variants having the same directions of effects across different SUDs as SUD-shared and developed a pipeline to prioritize SUD-shared genes. In total, 220 loci were identified with 40 novel loci not reported as SUD-associated. We prioritized 785 SUD-shared genes and identified 183 FDA approved drugs targeting these genes. By using a large real-world data from Optum® Clinformatics®, 7 drugs showed significantly reduced hazard ratio (HR) to develop SUDs: Topiramate (HR=0.44, 95% confidence interval (CI): 0.42 -0.47), Aripiprazole or Cariprazine (HR=0.88, 95% CI: 0.78-0.88), Desipramine, Imipramine, or Nortriptyline (HR=0.89, 95% CI: 0.84-0.94), Methylphenidate (HR=0.84, 95% CI: 0.78-0.91), demonstrating that they may be repurposed to treat SUDs.

Keywords: Drug repurposing, genome-wide association studies, substance use disorders, substance use disorders-shared genes, real-world data analysis.

Title: An Informatics Bridge to Improve the Design and Efficiency of Phase I Clinical

Trials for Anticancer Drug Combinations

Author list: Lei Wang

Detailed Affiliations:

The Ohio State University

Abstract: Prior preclinical and clinical knowledge is critical for designing effective and efficient cancer drug combinatory trials. In this study, we summarized critical databases of drug combination toxicity and pharmacokinetics. We further conducted a feasibility and utility study that demonstrates how different data sources can contribute to and assist phase I trial designs. Single-drug and drug combination toxicity and pharmacokinetic data were primarily reviewed from several databases. We focused on the MTD, dose-limiting toxicity (DLT), toxicity, and pharmacokinetic profiles. To demonstrate the feasibility and utility of these data sources in improving trial designs, phase I studies reported in ClincalTrials.gov from January 1, 2018 to December 31, 2018 were used as examples. We evaluated whether and how these studies could have been designed differently given toxicity and pharmacokinetic data. None of the existing pharmacokinetic and toxicity databases contain either MTD or DLT. Among 268 candidate trials, four drug combinations were studied in other phase I trials before 2018; 185 combinations had complete or partial information on drug interactions or overlapping toxicity, and 79 combinations did not have available information. Two drug combination trials were selected as case studies. The nivolumab-axitinib trial could have been designed as a dose deescalating study, and the vinorelbine-trastuzumab emtansine trial could have been designed with a lower dose of either drug. Public data sources contain significant knowledge of the drug combination phase I trial design. Some important data (MTD and DLT) are not available in existing databases but in the literature. Some phase I studies could have been designed more efficiently with additional preliminary data.

Keywords: Phase I clinical trial; cancer; drug combination; knowledge base

Title: Building an explainable graph neural network by sparse learning for the drug-protein binding prediction

Author list: Yijie Wang

Detailed Affiliations:

Indiana University

Abstract: Explainable Graph Neural Networks have been developed and applied to drug-protein binding prediction to identify the key chemical structures in a drug that have active interactions with the target proteins. However, the key structures identified by the current explainable GNN models are typically chemically invalid. Furthermore, a threshold must be manually selected to pinpoint the key structures from the rest. To overcome the limitations of the current explainable GNN models, we propose SLGNN, which stands for using Sparse Learning to Graph Neural

Networks. It relies on using a chemical-substructure-based graph to represent a drug molecule. Furthermore, SLGNN incorporates generalized fused lasso with message-passing algorithms to identify connected subgraphs that are critical for the drug-protein binding prediction. Due to the use of the chemical-substructure-based graph, it is guaranteed that any subgraphs in a drug identified by SLGNN are chemically valid structures. These structures can be further interpreted as the key chemical structures for the drug to bind to the target protein. We test SLGNN and the state-of-the-art competing methods on three real-world drug-protein binding datasets. We have demonstrated that the key structures identified by our SLGNN are chemically valid and have more predictive power.
Keywords: Graph Neural Network, Interpretable model, Sparse learning, Drug-protein binding prediction

Title: Combining genetics and real-world patient data fuel ancestry-specific target and drug discovery in Alzheimer's disease
Author list: Yuan Hou
Detailed Affiliations:
Cleveland Clinic
Abstract: Although high-throughput DNA/RNA sequencing technologies have generated massive genetic and genomic data in human disease, translation of these findings into new patient treatment has not materialized. Method: To address this problem, we have used Mendelian randomization (MR) and large patient's genetic and functional genomic data to evaluate druggable targets using Alzheimer's disease (AD) as a prototypical example. We utilized the genetic instruments from 6 celltype specific eQTLs and tested the outcome of MR independently across 7 genome-wide association studies (GWAS). Results: We identified 25 drug targets for AD. We pinpointed that the inflammatory target of epoxide hydrolase 2 (EPHX2) emerged as a potent AD target in EAs, and treatment of AD transgenic rats with an EPHX2 inhibitor was therapeutic. We also identified 23 candidate drugs associated with reduced risk of AD in mild cognitive impairment (MCI) patients after analysis of ~80 million electronic health records. Using a propensity score-matched design, we identified that usage of either apixaban (hazard ratio [HR] = 0.74, 95% confidence interval [CI] 0.69 – 0.80) and amlodipine (HR = 0.91, 95% CI 0.88 – 0.94) were both significantly associated with reduced progression to AD in people with MCI. Conclusion: In summary, combining genetics and real-world patient data identified ancestry-specific therapeutic targets and medicines for AD and other neurodegenerative diseases if broadly applied.
Keywords: Mendelian randomization, AD, GWAS, EPHX2, drug, target

Title: Identifying repurposable treatments in patient subpopulations.
Author list: Pengyue Zhang
Detailed Affiliations:
Indiana University
Abstract: Real-world data mining has the potential to identify precise relationships between drug responses and patient characteristics. We investigated drug responses in Alzheimer's disease (AD) with a special awareness on patient characteristics. In a multidisciplinary study, we observed both real-world evidence and genetic associations supporting telmisartan as a candidate repurposable drug for AD in African Americans. Additionally, we identified candidate repurposable drugs for AD in patients with neuroinflammation-related conditions.
Keywords: Alzheimer's disease, drug repurposing, neuroinflammation, real-world data, subpopulation

Title: So You Think You've Found a Target? Computational Simulation Methods for Hit Identification
Author list: Michael Robo
Detailed Affiliations:
Molecular Innovation, Indiana Biosciences Research Institute, Indianapolis, IN, USA
Abstract: This talk is intended to help bioinformaticians and other non-specialists understand the computational hit generation process for small molecule drug development. We will discuss the key questions to ask to determine how suitable a target is for small molecule drug development. We will then review some of the commonly used methods for structure-based virtual screening, including molecular docking, molecular dynamics, alchemical free energy calculations, and machine-learning strategies. We will conclude with a case study on a successful screen identifying novel type II inhibitors of the LYN kinase.
Keywords: Virtual Screening, Hit Identification, Molecular Simulation

Workshop – Bioinformatics Meet Biosignals: Opportunities and Challenges

August 4th
9:20 AM – 12:20 PM
Room: 350

Chairs: Haoqi Sun, Chen Huang

Title: Bioinformatics Meets Biosignals: Opportunities and Challenges

Author list: Haoqi Sun[1]
Detailed Affiliations:
[1]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
Abstract: Biosignals, such as brain wave (electroencephalography, EEG) during sleep, contain rich information about the function and health. However, the intersection between biosignals and bioinformatics is understudied due to the lack of interdisciplinary collaborations. Here, I describe opportunities and challenges in this field, focusing on the molecular basis of sleep electrophysiology as measured by EEG microstructures, as well as their implications on brain health.
Keywords: Biosignal, omics, sleep, electroencephalography

Title: Leveraging Clinical Biobanks and Genetics to Understand Sleep Apnea and Related Comorbidities
Author list: Brian Cade[1]
Detailed Affiliations:
[1]Department of Medicine, Brigham Women's Hospital, Harvard Medical School, Boston, MA, USA
Abstract: We have recently identified associations between sleep apnea (OSA) and hundreds of diseases in a large clinical biobank. Validation of these associations using sleep clinic data, identifying shared genetic architecture, and organizing prioritized comorbidities into multimorbidity clusters are important next steps to improve our understanding of sleep apnea and its consequences. In this proposed talk, I will describe our approach to phenotyping thousands of clinical sleep recordings using advanced polysomnographic traits (endotypes and burdens) that have increased associations with comorbidities compared to traditional measures of sleep apnea. Multivariate analyses indicate that no single polysomnographic trait best captures these case-control associations. We have started to measure the heritability and genetic correlations of OSA endotypes and burdens and anticipate performing multivariate genome-wide association studies to improve study power. Finally, I will describe recent analyses to group associated comorbidities into age-dependent topic models.
Keywords: Genomics, genome-wide association studies, obstructive sleep apnea, endotype

Title: Sleep Architecture Biomarkers of Psychiatric Disease
Author list: Shaun Purcell[1]
Detailed Affiliations:
[1]Department of Psychiatry, Brigham Women's Hospital, Harvard Medical School, Boston, MA, USA
Abstract: Sleep can now be measured using increasingly scalable and non-invasive sensor technologies, making it an attractive potential source of future novel objective biomarkers, with applications across a range of physical and mental conditions. Here I outline applications to psychiatric disease (primarily schizophrenia) and cognitive aging, and discuss some of the challenges faced when developing biomarkers based on sleep biosignals.
Keywords: sleep, psychiatric disease, biosignal

Title: Reimagining Sleep Medicine using AI-based Physiology-guided Digital Twins
Author list: Ankit Parekh[1]
Detailed Affiliations:
[1] Icahn School of Medicine at Mount Sinai, New York, NY, USA
Abstract: Despite affecting over a billion people worldwide, sleep disorders such as obstructive sleep apnea (OSA) remain poorly characterized in terms of symptom severity, treatment response, and long-term health consequences. Conventional metrics like the Apnea-Hypopnea Index (AHI) fall short in capturing the true physiological burden of these conditions and offer limited insight into patient heterogeneity. In this talk, I will present a transformative framework for sleep medicine—anchored in the development of physiology-guided, AI-based digital twins. These virtual representations of patients integrate time-series data from sleep studies (e.g., EEG, airflow, oxygen saturation) with multimodal clinical inputs to simulate disease trajectories, predict outcomes, and personalize therapy.
Keywords: sleep, digital twin, artificial intelligence

Title: Multi-omics in Neurodegenerative Diseases
Author list: Bruno Benitez[1]
Detailed Affiliations:
[1]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
Abstract: Integrating multiple omics by aligning genomic, proteomic, and metabolomic data provides a comprehensive view of biological and pathological processes. Genome sequencing enables a comprehensive analysis of the entire genome, while bulk and single-cell transcriptomics provide an extensive view of the transcriptome. Meanwhile, advancements in proteomics and metabolomics have not kept up with this technology. There is a need for a statistical framework or bioinformatics tools to enable the seamless integration of these large datasets. Artificial intelligence is beginning to bridge these gaps. The application of cross-omics to neurodegenerative diseases has allowed us to stratify patients based on their molecular landscape.
Keywords: Multi-omics, neurodegenerative diseases, artificial intelligence

Title: Y-chromosome loss in cancer: single-cell insights into origins and consequences
Author list: Shiwei Yin[1], Yusi Fu[1], Jun Xia[1]
Detailed Affiliations:
[1] Institute of Biosciences and Technology, Texas A&M University, Houston, TX, 77030, USA
Abstract: Mosaic loss of the Y chromosome (LOY) is observed in cancer, aging, and cardiomyopathy; however, its origins, mechanisms, and functional impact remain insufficiently understood. Bulk sequencing has limited sensitivity for capturing copy-number heterogeneity, leaving the landscape of LOY in clinical specimens unresolved. To address this gap, we applied an ultra-sensitive single-cell DNA copy-number method—capable of detecting copy-number alterations as low as 10 kb resolution—to thousands of cells from multiple gastrointestinal (GI) cancers. Our analysis quantified LOY frequency across major GI tumor types and determined whether LOY could be detected in peripheral blood mononuclear cells (PBMCs) or arose solely as a somatic event within tumors. We further characterized the extent of genomic instability in LOY clones, revealing frequent co-occurrence of additional genomic

events. Integrating single-cell RNA-seq data showed that LOY correlates with altered immune cell compositions and key immune-response genes, suggesting a potential role in modulating sensitivity to immunotherapy. In parallel, longitudinal culture of GI cancer and precancer cell lines demonstrated the progressive expansion of LOY subclones, shedding light on the dynamics of Y-chromosome loss over time. Collectively, these single-cell genomic analyses offer a high-resolution view of copy-number evolution in GI cancers, clarify the sources and consequences of LOY, and identify potential biomarkers and therapeutic targets associated with Y-chromosome loss.
Keywords: LOY, single-cell, copy-number, gastrointestinal, neoantigen, immunotherapy

Title: Computational Techniques for Deciphering Cancer Genomics and the Tumor Microenvironment at Single-Cell Resolution
Author list: Jinzhuang Dou[1]
Detailed Affiliations:
[1]Biomedical Informatics and Data Science, The University of Alabama at Birmingham, AL, USA
Abstract: Our understanding of cancer gene mutations and how cancer evades the immune system has led to the development of targeted therapies and immunotherapies. Single-cell sequencing further enhances these treatments by revealing tumor genetic heterogeneity and elucidating their interactions with immune cells. In this talk, I will present our computational efforts to advance cancer research at single-cell resolution. First, I will introduce Monopogen, a computational tool for single-nucleotide variant (SNV) calling in single-cell sequencing data. Leveraging Monopogen maximizes the genetic information from available single-cell sequencing data, leading to immediate benefits in genetic ancestry mapping and somatic clonal lineage delineation. Second, I will demonstrate a novel mathematical solution, bi-order canonical correlation analysis (bi-CCA), which iteratively aligns rows and columns between data matrices. Bi-CCA effectively integrates two distinct single-cell modalities derived from the same sample. Through bi-CCA, we deepen our understanding of immune cell therapy processes from a comprehensive multi-omic perspective. Looking forward, these computational tools hold great promise for uncovering new insights and improving personalized cancer treatments.
Keywords: Cancer evolution, Immunotherapy, Single cell, Ancestry, Multi-omics

Title: Distinct Signatures of Tumor-Associated Macrophages in Shaping Immune Microenvironment and Patient Prognosis

Author list: Chongming Jiang[1]

Detailed Affiliations:
[1] Terasaki Institute for Biomedical Innovation, Los Angeles, CA 90024, USA

Abstract: Background: Renal cell carcinoma (RCC) comprises 90% of adult kidney cancers, characterized by significant heterogeneity within its tumor microenvironment. This study tests the hypothesis that tumor-associated macrophages (TAMs) influence RCC progression and treatment responses. Using immunomics, we investigated the prognostic value of TAM signatures in the RCC tumor immune microenvironment (TIME). **Methods**: Single-cell RNA sequencing data from RCC patients were analyzed to develop eight distinct TAM signatures. A machine learning model predicting patient survival was built using TCGA data and validated across multiple independent RCC cohorts. Model performance was evaluated using Kaplan-Meier survival analysis, receiver operating characteristic (ROC) curves, principal component analysis

(PCA) visualization. **Results**: We identified diverse TAM subpopulations within the RCC TIME, highlighting significant prognostic implications. Specific TAM signatures correlated strongly with patient survival, macrophage infiltration, and known TAM markers. A TAM risk model effectively stratified patients into distinct risk categories, with the low-risk group showing significantly improved overall survival. **Conclusions**: Our findings clarify the complex roles of TAMs within RCC and their impact on patient prognosis. The established TAM risk model offers valuable prognostic markers and identifies potential therapeutic targets to enhance RCC treatment efficacy.

Keywords: tumor-associated macrophages, renal cell carcinoma, prognosis; tumor immune microenvironment, machine learning

Workshop – AI and Applications for Better Understanding Disease Mechanisms

August 4th
2:20 PM – 5:20 PM
Room: 320

Chairs: Xubo Song

Title: Evaluate, standardize, and optimization bioinformatics software documentation using AI-agents

Author list: Shaopeng Gu[1], Cankun Wang[1,2], Shaohong Feng[1], Qin Ma[1,2], Anjun Ma[1,2,*]
Detailed Affiliations:
[1]Department of Biomedical Informatics, Columbus, OH, USA 43210; [2] Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center – The James, Columbus, OH, USA 43210

Abstract: The exponential growth of omics data and the proliferation of computational tools have transformed biomedical research, enabling the exploration of complex biological systems at unprecedented resolution. Thousands of tools now exist for single-cell RNA sequencing, spatial transcriptomics, and other omics technologies, many powered by advanced artificial intelligence (AI) and deep learning methods. However, significant challenges remain in usability, reproducibility, and scalability—often stemming from inconsistent and non-standardized software documentation. These issues limit access for non-specialist users, reduce community engagement, and hinder reproducibility. While large language models (LLMs) are beginning to show promise in multi-omics data analysis, their potential for improving software engineering and documentation remains largely underexplored. To address this gap, we developed BioGuider, an innovative LLM-powered platform designed to standardize and enhance documentation during omics tool development. BioGuider operates through multiple specialized AI agents that (1) parse and analyze source code and documentation to identify logical structure, (2) evaluate documentation quality and reproducibility using an in-house scoring framework, and (3) automatically revise and generate key documentation components—including in-line code comments, user tutorials, vignettes, README files, and user guides. BioGuider is the first chat-based

assistant specifically built for bioinformatics tool development, establishing a new standard for documentation generation. These standards quantitatively evaluate the readability, documentation density, code-documentation ratio, structure quality, content quality, example coverage. Beyond assessing existing packages, we envision BioGuider as a proactive guide that helps developers meet publication and journal documentation standards prior to manuscript submission.
Keywords: Software development, quality control, documentation standardization, large language model

Workshop – Integrative genomics and epigenomics to link GWAS variants to function

August 4th
2:20 PM – 5:20 PM
Room: 350

Chairs: Hongbo Liu, Kaixiong Ye

Title: Precision Nephrology: The Role of Genetics in Kidney Health
Author list: Atlas Khan
Detailed Affiliations:
Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY, USA
Abstract: Chronic kidney disease (CKD) affects an estimated 10-13% of the global population and represents a major cause of morbidity, premature mortality, and healthcare burden worldwide. The pathogenesis of CKD is influenced by a complex interplay of genetic and environmental factors. Over the past decade, genome-wide association studies (GWAS) have identified hundreds of common variants associated with estimated glomerular filtration rate (eGFR), a central biomarker for kidney function. These discoveries have led to the development of genome-wide polygenic scores (GPS) that aggregate the effects of common variants to quantify an individual's genetic predisposition to CKD. In our recent work, we developed and validated a CKD GPS across diverse ancestral populations, demonstrating its utility for population-level risk stratification for CKD (**Khan et al. Nature Medicine 2022**).
While GPS provides a valuable tool for risk stratification in the general population, it fails to capture rare, high-penetrance protein-coding variants that underlie monogenic kidney diseases. Notably, disorders such as autosomal dominant polycystic kidney disease (ADPKD) and COL4A-related nephropathies, caused by pathogenic variants in *PKD1*, *PKD2*, and *COL4A3-5*, account for a significant portion of early-onset and progressive CKD. To address this gap, we leveraged large-scale exome sequencing data from population-based cohorts, including the UK Biobank and All of Us Research Program, to study the combined impact of polygenic and monogenic variation on kidney disease risk (**Khan et al., Nature Communications, 2022**).
In this talk, I will present a comprehensive risk prediction framework that integrates common and rare genetic variation to more accurately assess CKD risk. Our approach enhances the precision of risk prediction, particularly for individuals with an intermediate GPS who also carry a high-risk monogenic variant, and might improve our ability to detect early kidney dysfunction across diverse ancestral backgrounds.
Keywords: CKD, GWAS, PRS

Title: Unraveling the Molecular Heterogeneity of Severe Acute Malnutrition: Multi-omic Insights

Author list: Natasha C. Lie[1,2], Yixing Han[2], Qing Li[2], Aarti Jajoo[2], Aparna Haldipur[2], Emilyn Banfield[2], Shanker Swaminathan[3], Sharon Howell[4], Orgen Brown[4], Roa Sadat[3], Nancy J. Hall[3], Kwesi Marshall[4], Katharina V. Schulze[3], Thaddaeus May[5], Marvin E. Reid[4], Carolyn Taylor-Bryan[4], Mark J. Manary[6,7,8], Indi Trehan[7,9], Mamana Mbiyavanga[10], Wisdom A. Akurugu[10], Colin A. McKenzie[4^], Dhriti Sengupta[11], Elizabeth G. Atkinson[3,12], Ananyo Choudhury[11], Neil A. Hanchard[1,2,3,6,*]

Detailed Affiliations:

[1] Graduate Program in Integrative Molecular and Biomedical Sciences, Baylor College of Medicine, Houston, TX, USA. [2] Childhood Complex Disease Genomics Section, National Human Genome Research Institute, National Institutes of Health, Bethesda, USA. [3] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [4] Tropical Metabolism Research Unit, Caribbean Institute for Health Research, University of the West Indies, Mona, Jamaica. [5] Department of Internal Medicine, Baylor College of Medicine, Houston, TX, USA. [6] USDA/ARS/Children's Nutrition Research Center, Baylor College of Medicine, Houston, TX, USA. [7] Departments of Paediatrics and Child Health and Community Health, Kamuzu University of Health Sciences, Blantyre, Malawi. [8] Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, USA. [9] Departments of Pediatrics, Global Health, and Epidemiology, University of Washington, Seattle, WA, USA. [10] Computational Biology Group, Faculty of Health Sciences, University of Cape Town, Western Cape, South Africa. [11] Sydney Brenner Institute for Molecular Bioscience (SBIMB), University of the Witwatersrand, Johannesburg, South Africa. [12]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA

Abstract: Severe acute malnutrition (SAM) remains a global health emergency, directly or indirectly responsible for over 400,000 childhood deaths each year. Among its clinical forms, edematous SAM (ESAM, including kwashiorkor and marasmic-kwashiorkor) is more lethal than non-edematous SAM (NESAM or marasmus), despite comparable nutritional deficiencies and environments. ESAM is also the predominant form in populations from East Africa and the Caribbean. However, the molecular mechanisms underlying why some children get ESAM and others NESAM remain poorly understood.

To address this gap, we have developed a multi-omic framework to interrogate clinically phenotyped cohorts from Jamaica and Malawi. Our initial analyses included genome-wide DNA methylation profiling of buccal cells from 309 children revealed widespread hypomethylation in ESAM cases, particularly at genes implicated in metabolic and liver-related pathways. These epigenetic alterations were absent in adults recovered from SAM, pointing to disease-specific, dynamic methylation changes potentially driven by disrupted OCM. In parallel, intracontinental admixture mapping and targeted genotyping of 103 genes involved in one-carbon metabolism (OCM) across 711 children. We identified seven loci—such as *MTHFR*, *PRICKLE2*, and *PLD2*— with evidence of significant association with ESAM. These loci were enriched on East African ancestral haplotypes and located within genomic regions under recent positive selection, suggesting a potential evolutionary influence on disease susceptibility.

To deepen our understanding of SAM heterogeneity, we are integrating whole-genome sequencing, transcriptomics, methylation, and metabolomics data from ESAM and NESAM patients on the NHGRI AnVIL cloud platform. Scalable, machine learning–enabled pipelines will allow for population-aware variant calling, methylation quantitative trait loci (meQTL) mapping, functional enrichment, and network-based analyses. This integrative approach aims to elucidate genotype–epigenotype–phenotype relationships and identify molecular networks contributing to ESAM, with a strong emphasis on reproducibility, equity, and cross-cohort validation. Together, our findings support a model in which inherited variation in OCM pathways interacts with environmental and epigenetic factors to drive ESAM pathogenesis. By revealing molecular signatures unique to high-risk populations, our study lays the foundation for precision nutrition strategies and demonstrates the potential of ancestry-aware, cloud-based multi-omic research in addressing hidden drivers of pediatric disease.
Keywords: kwashiorkor; marasmus; multi-omics; genetic variants, admixture; local ancestry

Title: Leveraging chromatin accessibility data to understand complex traits
Author list: Liyang Yu[1], Siming Zhao[1]
Detailed Affiliations:
[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College
Dartmouth Cancer Center
Abstract: Thousands of significant signals have been identified from Genome-Wide Association Studies (GWAS). As many associated variants locate in regulatory regions, cell type or tissue specific regulatory elements offer valuable insights in delineating disease related cell types or tissues. The development of scATAC-seq offers unprecedented resolution of different cell states defined by their chromatin accessibility profiles. Open chromatin regions often indicate regulatory elements; the enrichment of GWAS variants in open chromatin regions of a particular cell state indicates that this cell state is associated with the GWAS trait. Most current efforts identify disease associated cell types using bulk ATAC-seq data or construct pseudo bulk data for cell types identified from scATAC-seq; a disease relevance score can be derived on a cell-type level. However, such analyses neglect heterogeneity within a cell type, and for some scATAC-seq data, the cell states can be continuous rather than discrete, leading to ambiguity in cell typing. In this talk, I will present a new method to assess the disease relevance at cell level leveraging single cell ATAC-seq data. Our method leverages the polygenic signals of disease variants in GWAS data to assess its enrichment over the background at cell level. We overcome the sparsity issue of single cell ATAC-seq data through co-regulatory patterns of open chromatin regions across cells. Through simulations we found our method outperformed the states of art methods, providing more accurate cell level disease relevance scores and more effectively leverage single cell ATAC data to identify causal variants. We demonstrate the usefulness our method on single cell ATAC atlas data for a variety of complex traits.
Keywords: Chromatin accessibility, complex traits, statistical genetics, single cell ATAC-seq.

Title: Integrative genomics and epigenomics reveal functions of non-coding variants
Author list: Hongbo Liu[1,2,3,4] and Katalin Susztak[2,3,4]
Detailed Affiliations:
[1]Department of Biomedical Genetics, University of Rochester Medical Center, Rochester, NY 14642, USA. [2]Department of Medicine, Renal Electrolyte and Hypertension Division, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Institute of Diabetes Obesity and Metabolism, University

of Pennsylvania, Philadelphia, PA 19104, USA. [4]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA.

Abstract: Background: Genome-wide association studies (GWAS) have identified numerous DNA sequence variants associated with complex human diseases. However, over 90% of disease-associated variants reside in noncoding genome regions, and their functions in complex diseases remain largely unknown—a problem often referred to as the 'variant-to-function' problem. **Methods**: To link non-coding variants to functions in human diseases, we integrated various genomic and epigenomic datasets to identify the regulatory variants by developing several computational strategies, including methylation quantitative trait loci (meQTL), allele-specific expression (ASE), and allele-specific expression accessibility (ASA). In particular, we developed a statistical model, Open4Gene, to link non-coding variants to their target genes using single cell multiome data, which simultaneously profiles DNA accessibility and gene expression within the same cell. **Results**: We conducted a multi-ancestry GWAS mapping in 2.2 million individuals and identified over 1,000 independent loci associated with kidney function. Ancestry-specific analysis indicated an attenuation of newly identified signals on common variants in European ancestry populations and the power of population diversity for further discoveries. We defined genotype effects on allele-specific gene expression and regulatory circuitries in human kidneys and cells. We developed a statistical approach named Open4Gene, which identified 1,351 target genes of genetic variants located within open chromatin regions. By integrating these GWAS and multiome datasets (total 32 types), we found over 24,000 regulatory variants targeting more than 1,000 genes, with over 600 genes also targeted by coding variants. In particular, we discovered the convergence of coding and regulatory variants on 161 key disease genes, critical cell types (including proximal tubules), transcriptional regulators (including HNF4A), and potential drug targets for kidney disease, providing an integrative strategy for functional annotation of noncoding variants in complex human diseases.

Keywords: Human Genetics, epigenetics, kidney disease, non-coding variants, single-cell multiome

Title: Mechanistic annotation of GWAS loci for circulating fatty acids by single-cell omics and CRISPR screens

Author list: Huifang Xu[1], Haifeng Zhang[1], Ge Yu[1], Yitang Sun[1], Elijah Sterling[1,2], Saurav Choudhary[3], Pengpeng Bi[1], Kaixiong Ye[1,3]

Detailed Affiliations:

[1]Department of Genetics, University of Georgia, Athens, Georgia, USA; [2]Regenerative Bioscience Center, University of Georgia, [3]Institute of Bioinformatics, University of Georgia, Athens, Georgia, USA

Abstract: Fatty acids (FA) play crucial roles in human health, influencing the risk of developing various conditions, such as cardiovascular disease and dementia. While previous genome-wide association studies (GWAS) have identified hundreds of genetic loci associated with the circulating FA levels, the underlying biological mechanism linking these identified loci to FA metabolism remains largely unclear. Here, we integrate GWAS with single-cell multi-omics and single-cell CRISPR screens to systematically uncover the cellular and molecular mechanisms underlying FA-associated genetic loci. We colocalized GWAS signals for 19 FA traits with six types of multi-omics quantitative trait loci (QTL), including gene expression, protein abundance, DNA methylation, splicing, histone modification, and chromatin accessibility, to identify intermediate molecular phenotypes that mediate the associations between the genetic loci and 19 FA traits. We found that 35% of GWAS loci overlapped with QTL signals for at least one molecular phenotype. Notably, a locus (around genes *GSTT1/2/2B*) associated with total fatty acids, the percentage of omega-6 polyunsaturated fatty acids (PUFA) in total FAs, and total monounsaturated

fatty acids overlapped with QTL signals across all six molecular phenotypes. We analyzed single-cell RNA-seq data of over 100,000 cells from liver tissue to explore the cellular context. We discovered that hepatocyte cell populations, particularly those located in the periportal region, are enriched for genes associated with FA traits. To explore the regulatory function of FA-associated loci, we conducted a single-cell CRISPR screen in over 200,000 HepG2 liver cells, targeting 360 candidate regulatory elements (CREs) from fine-mapped FA trait variants. We identified target genes in cis for 298 CREs, providing a direct map of the regulatory relationship. Our integrative analysis reveals the molecular and cellular mechanisms regulating circulating fatty acid levels, providing mechanistic insights into the genetic architecture of fatty acid metabolism.
Keywords: Fatty acids, GWAS, single-cell multi-omics, single-cell CRISPR screen, Mechanistic annotation

Title: Linking Rare Non-Coding Regulatory Variants Associated with Human Longevity to Cellular Senescence via Integrative Functional Genomic Approaches
Author list: Jiping Yang[1]; HyeRim Han[1]; Jhih-Rong Lin[2]; Zhengdong Zhang[2]; Sofiya Milman[2]; Nir Barzilai[2]; Yousin Suh[1]
Detailed Affiliations:
[1] Department of Obstetrics and Gynecology, Columbia University Medical Center, New York, USA; [2] Department of Genetics, Albert Einstein College of Medicine, Bronx, New York, USA
Abstract: Centenarians, despite representing a tiny proportion of the global population, hold the key to access longevity. By decoding the genomes in a unique Ashkenazi Jewish (AJ) centenarian cohort, we have identified rare coding variants protective against age-related diseases, along with numerous non-coding variants with unknown functions. Non-coding variants, once considered "Junk DNA", are now known to enrich in cis-regulatory elements (CREs) that control transcriptional activity. However, the functional interpretation of non-coding variants remains challenging due to incomplete knowledge of regulatory elements, their mechanisms of action, and the cellular states and processes in which they function, let alone the identification of truly causal variants and their target genes. To partially address this challenge, we employed phenotypic CRISPR screens to discover longevity-associated variant-residing CREs capable of modulating cellular senescence. We prioritized rare regulatory variants identified in our AJ centenarian cohort by mapping non-coding variants in linkage disequilibrium (LD) to potential CREs annotated by Cis-element Atlas (CATlas). Pooled activation (CRISPRa) or inhibition (CRISPRi) using CRE-targeting sgRNAs alleviated cellular senescence in human mesenchymal stromal cells compared to non-targeting sgRNAs. Sequencing-based sgRNA enrichment analysis in endpoint cells identified putative senescence-modulating CREs. Surprisingly, almost all these CREs were located in intergenic or intronic non-promoter regions. To further elucidate the role of these senescence-modulating CREs in transcriptional regulation, we conducted transcriptome-based single-cell CRISPR interference screens to identify their cis-regulated causal genes and trans-effect networks, leading to the discovery of novel genes driving cellular senescence and potential targets for extending human healthspan and lifespan.
Keywords: Functional genomics, rare non-coding variant, longevity, CRISPR screen

Title: Identification of replicative aging and inflammatory aging signatures via whole-genome CRISPRi screens and GWAS meta-analysis
Author list: Lingling Wu[1,2,3†], Xiang Zhu[4,5,6,9†], Yanxia Liu[1,10], Dehua Zhao[1,11], Betty Chentzu Yu[2,3], Zheng Wei[2,3], Xueqiu Lin[1,2,3*], Lei S. Qi[1,7,8*]

Detailed Affiliations:
[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA. [2]Computational Biology Program, Fred Hutchinson Cancer Center, Seattle, WA 19024, USA. [3]Translational Data Science IRC, Fred Hutchinson Cancer Center, Seattle WA 19024, USA. [4]Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA. [5]Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA. [6]Department of Statistics, Stanford University, Stanford, 94305, CA, USA. [7]Sarafan ChEM-H, Stanford University, Stanford, CA 94305, USA [8]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [9]Currrent address: Calico Life Sciences, South San Francisco, CA 94080, USA. [10]Currrent address: Epicrispr Biotechnologies, South San Francisco, CA, USA. [11]Currrent address: Genomic Sciences, GSK, San Francisco, CA, USA
[†]These authors contributed equally to this work.
Abstract: Aging is a major risk factor for chronic diseases and cancer. Cellular aging, particularly in adult stem cells, offers a high-throughput framework for dissecting the molecular mechanisms of aging. We performed multiple genome-wide CRISPR interference (CRISPRi) screenings in human primary mesenchymal stem cells (MSC) derived from adipose tissue during either replicative senescence or inflammation-induced senescence. These screens revealed distinct sets of potential novel regulators specific to each senescence pathway. Combining our perturbation-based functional genomic data with 405 genome-wide association study (GWAS) datasets, including 50 aging-related studies, we found that the inflammatory aging signatures identified from CRISPRi screenings were significantly associated with diverse aging processes, suggesting novel molecular signatures for analyzing and predicting aging status and aging-related disease. The signatures verified through comprehensive functional genomics and genetic analyses may provide new targets for modulating the aging process and enhancing the quality of cell therapy products.
Keywords: Inflammatory aging, CRISPR-screening, GWAS

Workshop – Data science solutions for spatial transcriptomics

August 5th
9:20 AM – 12:20 PM
Room: 320

Chairs: Travis Johnson

Title: SpatialGE: An Interactive Web Platform for Accessible and Reproducible Spatial Transcriptomics Analysis

Author list: Xiaoqing Yu

Detailed Affiliations:
Moffit Cancer Center

Abstract: Spatial transcriptomics (ST) enables the study of gene expression in spatial context, offering insights into the tumor microenvironment. However, analyzing ST data remains a challenge for non-

computational researchers. We developed *spatialGE*, a point-and-click web application built on the spatialGE R package to make ST analysis accessible, reproducible, and interactive. The platform supports domain detection (SpaGCN, MILWRM), cell deconvolution (STdeconvolve), and phenotyping (InSituType), with workflows for both Visium and single-cell ST data such as CosMx and Xenium. Users can perform multi-step analyses, compare results across samples, and overlay results on tissue images. Reproducibility is ensured through parameter tracking and downloadable outputs. We demonstrated the applications of *spatialGE* to melanoma brain metastases and Merkel cell carcinoma, highlighting how the platform enables hypothesis generation and biological insight without requiring programming experience.

Keywords: Spatial Transcriptomics, Deconvolution, Web applications

Title: Spatial Resolved Gene Regulatory Networks Analysis

Author list: Zhana Duren

Detailed Affiliations:
Indiana University School of Medicine

Abstract: Integrating spatial transcriptomics – which maps gene expression location within tissues – with single-cell multi-omics data, profiling gene expression and chromatin accessibility (or other epigenomic data), offers powerful insights into gene regulation. However, commercially available kits for simultaneous spatial multi-omics profiling are currently unavailable, hindering widespread data generation. Here, we present ISON (Integrated Spatial Omics Network), a novel computational method to infer spatial-resolved gene regulatory networks by leveraging existing single-cell multiome data and spatial transcriptomics data. ISON accurately predicts omics profiles for spatial spots and reconstructs spatially resolved gene regulatory networks, demonstrating scalability in both time and memory. Importantly, ISON omics prediction preserves cis- and trans- regulatory information and enables estimation of transcription factor (TF) activity at the spot level, distinguishing between TFs even within the same family – a capability absents in approaches relying solely on ATAC-seq data. Application of ISON to Alzheimer's disease data reveals disease- and age-specific spatial gene regulatory modules, highlighting its potential for uncovering spatially organized mechanisms driving complex biological processes.

Keywords: Single cell, Multiomics, Gene regulatory networks, Spatial omics

Title: Identifying Key Regulators of Amyloid Beta Clearance from Single Cell Spatial Transcriptomics using Generalized Linear Mixed Effect Models

Author list: Debolina Chatterjee

Detailed Affiliations:
Indiana University School of Medicine

Abstract: Single-cell spatial transcriptomics (ST) enables the simultaneous profiling of gene expression and spatial organization within tissues, offering unprecedented insights into cellular microenvironments. To harness the full potential of such data, we present TAWGLE (Topology AWare Generalized Linear mixed Effect model), a novel statistical framework that integrates spatial topology, cell type identity, intercellular interactions, and disease context to dissect gene expression patterns. Focusing on Alzheimer's disease, where genome-wide association studies have highlighted *INPP5D* as a critical

regulator of microglial activity and amyloid beta (Aβ) accumulation, we applied our approach to three mouse models: wild-type (B6), amyloid-bearing (5xFAD), and amyloid-haplodeficient (5xFAD:KO). Our analysis reveals that genes associated with Aβ pathology show spatially-resolved interactions between microglia and astrocytes. Notably, *PSEN1, GNAQ,* and *COX8A* were upregulated in astrocytes near microglia in 5xFAD, while *CACNA1D, COX6C,* and *COX4I1* were downregulated in 5xFAD:KO. In microglia near astrocytes, *APH1B.C, PLCB1,* and *COX4I1* were upregulated in 5xFAD, while *SLC39A11, SLC11A2, PSEN1, APP,* and *TUBB5* were upregulated in 5xFAD:KO. Thus, we explore cellular neighborhoods within brain tissue, and identify genes associated with enhanced Aβ clearance.

Keywords: Spatial transcriptomics, Single cell, multiomics, Alzheimer's disease, Linear mixed effect models

Title: Leveraging Spatial Transcriptomics of Brain Tissue in Neurological Diseases

Author list: Oscar Harari

Detailed Affiliations:
The Ohio State University

Abstract: Single-cell omics approaches have revolutionized the profiling of brain cell molecular composition with remarkable detail. However, these methods often lose cellular context during tissue processing. Spatial transcriptomics offers the possibility of in situ profiling, providing crucial information about the cell neighborhood and, when combined with immunohistochemistry, relating cellular states to neuropathological lesions. Our research focuses on leveraging single-cell and spatial omics to profile both affected and healthy brain tissue, aiming to reveal cell-type specific changes associated with the etiology and progression of neurological diseases. In this presentation, we will explore how spatial transcriptomics can be employed to uncover novel insights into the pathogenesis of various neurological disorders. Alzheimer's disease presents as a heterogeneous disorder marked by diverse molecular mechanisms. Given the intricate nature of AD pathology, the relationships between cellular components and their spatial context is critical to understanding disease mechanisms. We employed the Visium HD platform to investigate dorsolateral prefrontal cortex tissues from late-stage AD patients, seeking to reveal spatial gene expression patterns in affected and unaffected regions at single-cell resolution. Immunofluorescence staining detected amyloid-beta and phosphorylated tau accumulations. This approach enabled us to examine the correlation between AD hallmarks and cellular gene expression profiles. Our findings indicate significant differences among cells proximal to amyloid plaques compared to the surrounding unaffected tissues, even among cell types with low representation like microglial and vascular populations.
Following ischemic stroke, pan-necrosis occurs at the injury core. Selective neuronal death is frequently observed in surrounding regions, but the molecular determinants underlying differential neuronal vulnerability remain unclear. To address this, we investigated whether homeostatic molecular pathways could predict the susceptibility or resilience of select neuronal populations to ischemia. Single-nuclei and spatial RNA sequencing were performed on the peri-infarct region of mice 24 hours post-tMCAO, the corresponding contralateral region, and sham mice to identify selectively vulnerable or resilient neurons. We identified genes expressed under homeostatic conditions in sham that predicted selective neuronal resilience or vulnerability. Utilizing the Vizgen MERSCOPE assay, we generated spatial maps, enabling observation of unique glial cell distribution within the infarct as well as spatial distribution of resilient and vulnerable cells within the peri-infarct and contralateral hemisphere.

In conclusion, the integration of single cell, spatially resolved transcriptomics, and immunochemistry significantly enhances our understanding of neurological diseases by elucidating specific spatial cellular niches in which cells mediate disease risk and progression.

Keywords: Spatial transcriptomics, Single cell, Neurodegenerative disease, Alzheimer's disease,

Title: A Statistical Framework to Improve the Design of Spatial Transcriptomics Experiments

Author list: Dongjun Chung

Detailed Affiliations:
The Ohio State University

Abstract: High-throughput spatial transcriptomics has recently gained significant attention and it can capture high-dimensional gene expression profiles in tissue samples at or near single-cell level while retaining the spatial location of each sequencing unit. This new technology provides unprecedented opportunities for biomedical research and has recently gained significant attention from various fields such as cancer research, neuroscience, and developmental biology. To effectively analyze this new type of data, various statistical and computational methods for spatial transcriptomics data analysis have been developed in recent years. However, while some efforts have been made to improve the design of these studies, it is still significantly understudied how to optimize key experimental factors of these experiments. In this talk, I will discuss spaDesign, our novel statistical framework for the design of spatial transcriptomics experiments, which aims to address this critical need. spaDesign is a statistically rigorously designed framework that employs Poisson Gaussian process and Fisher-Gaussian kernel mixture. It can easily simulate a range of spatial transcriptomics data with various sequencing depths, effect sizes, and spatial patterns, which allows rigorous estimation of needed total sequencing depths to detect spatial domains based on spatial transcriptomics experiments. We will demonstrate the utility and power of spaDesign using 10X Visium data from the human brain and the chicken heart.

Keywords: Spatial transcriptomics, Statistical frameworks, Poisson gaussian processes, Fisher-gaussian kernal mixtures

Title: Integrative Modeling of Gene Expression and Histology via Cross-Modal Alignment and Multi-Scale Graph Inference

Author list: Chao Chen

Detailed Affiliations:
Stony Brook University

Abstract: Spatial transcriptomics (ST) offers a powerful way to map gene activity within tissues, providing crucial insights into cellular diversity and spatial organization. When combined with histology images, ST data opens new avenues for enhancing whole slide image (WSI) prediction tasks, such as disease diagnosis and outcome forecasting. However, existing approaches often struggle to align gene expression with tissue images due to spatial distortions and modality differences, and they typically overlook the complex relationships between distant tissue regions. In this talk, we present methods that address these challenges. First, we introduce a novel ranking-based alignment method that captures nuanced cross-modal relationships between gene and image features while maintaining robustness across scales. This is further stabilized using a teacher-student self-supervised learning strategy to handle the noise and sparsity in gene expression data. Second, we propose **MERGE** (Multi-faceted hiErarchical

gRaph for Gene Expressions), a graph-based approach that models interactions across tissue patches by clustering them based on both spatial proximity and morphology. Through a hierarchical graph neural network (GNN), MERGE enables both local and long-range tissue interactions to inform gene prediction. We also examine the impact of various data smoothing techniques in ST, advocating for biologically grounded, gene-aware smoothing methods to reduce technical artifacts. Across multiple public datasets, our methods significantly outperforms current methods in tasks including gene prediction, slide-level classification, and survival analysis, demonstrating the promise of advanced feature alignment and multi-scale graph modeling for spatially informed biomedical insights.

Keywords: Spatial transcriptomics, Histopathology, Whole slide images, Graph neural networks, Multimodal models

Title: Utilizing Deep Transfer Learning to Identify High Risk Subpopulations of Cells in Single Cell and Spatial Omics Data

Author list: Travis S. Johnson

Detailed Affiliations:
Indiana University School of Medicine

Abstract: Leveraging single-cell gene expression profiles can significantly improve our understanding of diseases by associating single cells with traits such as disease subtypes, prognosis, and drug response. Although previous efforts have linked single cell clusters and groups with these attributes, they have primarily focused on changes in cell proportions while overlooking transcriptional changes at the single cell level. To further unravel cell heterogeneity with clusters and reveal nuanced behaviors of cellular subtypes, it is essential to assess the disease associations of individual cells. Previous methods often fail to capture complex patterns that are only discernible through summarizing non-linear relationships across multiple genes. The Diagnostic Evidence GAuge of Single-cells/Spatial-transcriptomics (DEGAS) framework advances these efforts by aligning single cells and/or spatial transcriptomics regions with patients through a unified latent space using nonlinear transformations learnt from deep neural networks (DNNs). Here, we present DEGAS version 2 (DEGASv2), which has been updated with optimal transport based transfer learning and improved time-to-event functionality, more advanced model architecture, and improved model baseline evaluations. DEGASv2 achieves superior performance in analyzing single cell and spatial transcriptomics datasets, including Alzheimer's disease (AD), multiple myeloma (MM) and prostate cancer (PDAC). On the MM discovery dataset, DEGASv2 enabled us to discover cell types that exhibited different drug response patterns over various time frames and were validated with multi-omic data from a time series of single cells that we generated, demonstrating a dangerous subtype of cell and a new therapeutic target.

Keywords: Spatial transcriptomics, Single cell, Transfer learning, Multiple myeloma, Alzheimer's disease, Prostate cancer

Workshop – Computational Omics for Precision Medicine and Drug Discovery

August 5th
9:20 AM – 12:20 PM
Room: 301

Chairs: Bin Chen, Qian Li

Title: Protein Language Model ESM3 Enables Superior Prediction of Complex Variant Effects Across ClinVar and DMS Benchmarks
Author list: Chang Li[1] and Xiaoming Liu[1]
Detailed Affiliations:
[1]Department of Global, Environmental, and Genomic Sciences, College of Public Health, University of South Florida, Tampa, Florida, USA.
Abstract: Accurate prediction of variant pathogenicity, particularly for complex mutations beyond single nucleotide variants (SNVs), remains a major challenge in genomic medicine. Traditional protein language models (PLMs) like ESM2 and AlphaFold focus on sequence or structure alone, limiting their ability to fully assess functional disruptions. Here, we demonstrate that ESM3, a multimodal protein model integrating sequence, structure, and function via geometric attention mechanisms, substantially advances variant effect prediction. Without relying on multiple sequence alignments, ESM3 shows strong performance across a wide range of variant types, including non-frameshift insertions/deletions (InDels) and complex variants such as stop-gain/loss mutations.
Benchmarking on Deep Mutational Scanning (DMS) and novel ClinVar datasets, ESM3 achieves superior prediction accuracy compared to current state-of-the-art tools, particularly excelling at complex and non-canonical variants where other models falter. Through case studies on GABRB3 and IRF6, we demonstrate that ESM3's cross-modality divergence and entropy metrics provide unique mechanistic insights, distinguishing gain-of-function (GOF) and loss-of-function (LOF) variants and highlighting domain-specific functional vulnerabilities. Our findings establish ESM3's zero-shot potential for variant effect prediction, particularly for poorly characterized or structurally disruptive mutations, offering new avenues for variant interpretation and precision medicine.
Keywords: protein language model, pathogenicity prediction, SNV, InDel

Title: Massive labeled transcriptomics as a resource of transcriptome representation learning and drug discovery
Author list: Bin Chen[1,2,3]
Detailed Affiliations:
[1]Department of Pharmacology and Toxicology, Michigan State University, East Lansing, MI 48824, USA. [2]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA. [3]Department of Pediatrics and Human Development, Michigan State University, Grand Rapids, MI 49503, USA
Abstract: Gene Expression Omnibus (GEO), the largest repository of transcriptomics data, houses more than millions of gene expression profiles from 200,000 studies and has been extensively explored for research; however, its metadata is presented in unstructured text and often lacks consistency due to varying submission formats. Manual curation is time-intensive and error-prone, posing challenges for dataset integration and downstream analyses. We introduce a GPT-based AI model to automate and standardize GEO metadata annotation, significantly improving efficiency, accuracy, and consistency. We

establish a structured annotation framework, integrating domain-specific mega prompts and standardization protocols to ensure uniformity across including strain, genotype, disease, and treatment details. We present a comprehensive annotation dataset that encompasses >100K mouse and human samples each, along with their transcriptome profiles. We further develop benchmarks for the prediction of labels from gene expression profiles using state-of-the-art transcriptome embedding methods. By combining the large-scale transcriptome data and our drug discovery platforms OCTAD and GPS, we can predict the therapeutic potential for any drugs or compounds against hundreds of diseases. We expect this dataset will become an essential resource of learning transcriptome and large-scale drug discovery.
Keywords: Large Language Models (LLMs); GPT-based Annotation; Drug Repositioning

Title: Generative AI for Human Genetics and Functional Genomics
Author list: Xinghua Shi[1]
Detailed Affiliations:
[1]Department of Computer & Information Sciences, Temple University, Philadelphia, PA 19122, USA
Abstract body: Generative artificial intelligence (GenAI) techniques have been addressing key challenges and introducing transformative opportunities in human genetics and functional genomics. In general, GenAI techniques that have been widely adopted in the field include variational autoencoders (VAEs), generative adversarial networks (GANs), large language models (LLMs), Transformers and diffusion models. Using our research products as examples, I will present the applications of these models to classical and emerging problems in genotype imputation, synthetic genotype generation, augmented chromatin characterization, and disease prediction.
Keywords: Generative Artificial Intelligence (GenAI), Functional Genomics, Large Language Models (LLMs), Computational Biology

Title: Distinct Mutational Profiles in Primary Sclerosing Cholangitis-Associated Cholangiocarcinoma Compared to *de novo* Cholangiocarcinoma
Author list: Shulan Tian[1], Filippo Pinto e Vairo[3], Ahmad H. Ali[2], Huihuang Yan[1], Bryan M. McCauley[4], Brian D. Juran[2], Tony C. Luehrs[4], Fan Leng[5], Cameron M. Callaghan[6], Jacob A. Frank[4], Sicotte Hugues[1], Sebastian M. Armasu[4], Robert A. Vierkant[4], Jan B. Egan[3], Zhifu Sun[1], Nicholas B. Larson[4], Eric W. Klee[1,3], Konstantinos N. Lazaridis[2,3]
Detailed Affiliations:
[1]Division of Computational Biology; [2]Division of Gastroenterology and Hepatology; [3]Center for Individualized Medicine and Department of Clinical Genomics; [4]Division of Clinical Trials and Biostatistics; [5]Data Analytics and Integration; [6]Department of Radiation Oncology, Mayo Clinic
Abstract: Introduction: Cholangiocarcinoma (CCA) is a rare and aggressive malignancy characterized by etiologic heterogeneity and poor survival. Primary sclerosing cholangitis (PSC) is the most recognized risk factor for CCA in Western countries. PSC-associated CCA (PSC-CCA) is a leading cause of morbidity and mortality in PSC patients and exhibits distinct clinical features compared to those in *de novo* CCA. However, the molecular mechanisms driving these two subtypes of CCA remain largely unexplored. This study aimed to characterize the spectrum and prevalence of germline genetic variants in pathologically confirmed PSC-CCA, and *de novo* CCA as well as PSC patients without CCA (PSC-w/o CCA). **Material and method:** This retrospective cross-sectional study included 301 patients with PSC-w/o CCA and 170 patients with CCA (PSC-CCA, n=88; *de novo* CCA, n=82) identified from two population genomics studies conducted at Mayo Clinic between 2016 and 2023. Their diagnoses, phenotypes, outcomes, as well as medical and family histories were obtained from electronic health

records (EHRs) and self-reported questionnaires. Exome sequencing of these patients was conducted with genomic DNA, and genetic variants were identified using bioinformatics workflow following the Genome Analysis Toolkit (GATK) best practices. A comprehensive list of c*ancer susceptibility genes* was compiled from prior cancer studies. Functional annotation and pathogenicity assessment of cancer-associated genetic variants were performed according to current American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines. **Result and discussion**: Analysis of exome sequencing data from 471 patients identified 53 pathogenic/likely pathogenic (P/LP) germline variants across 25 cancer susceptibility genes (CSGs) in 10.8% (51/471) of patients. The highest prevalence of P/LP germline variants was observed in PSC-CCA patients (13.6%, 12/88), followed by PSC-w/o CCA (10.0%, 31/301) and *de novo* CCA (9.76%, 8/82). Interestingly, PSC-CCA patients exhibited P/LP germline variants mainly in moderate-, low-penetrance, and/or autosomal recessive genes, with significant enrichment in the Fanconi anemia DNA repair pathway. In contrast, patients with *de novo* CCA predominantly carried P/LP germline variants in the tumor suppressor genes that are key players in homologous recombination repair pathway. Similarly, germline variants led to differentially altered metabolic and signal pathways observed between PSC-CCA and *de novo* CCA patients. **Conclusion:** These findings provide key insights into distinct CCA subtypes and call for an effort to systematically study germline testing of patients with PSC-CCA and *de novo* CCA as an approach to inform personalized approaches to screening, clinical management and targeted therapy of CCA in these patients.
Keywords: primary sclerosing cholangitis; cholangiocarcinoma; cancer susceptibility genes; exome sequencing; DNA repair

Title: High-resolution multi-omic dissociation of brain tumors with multimodal autoencoder
Author list: Jiao Sun[1], Ayesha Malik[2], Tong Lin[1], Ayla Bratton[2], Kyle Smith[3], Yue Pan[1], Arzu Onar-Thomas[1], Giles W. Robinson[4], Wei Zhang[2], Paul A. Northcott[3], Qian Li[1*]
Detailed Affiliations:
[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, 38105. [2]Department of Computer Science, University of Central Florida, Orlando, FL, 32816. [3]Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN, 38105. [4]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, 38105
Abstract: Single-cell technologies enable high-resolution profiling of molecular dynamics in developmental and cancer biology. But heterogeneity and complexity of tumors may hinder the lineage cell mapping in developmental origins or dissection of tumor microenvironment, requiring digital dissociation of bulk tissues. Many deconvolution methods focus on transcriptomic assay using scRNA-seq as reference, not easily applicable to other omics due to ambiguous cell markers and unexpected biological difference between reference and target tissues. Here, we present MODE, a multimodal autoencoder pipeline linking multi-dimensional molecular features to jointly predict personalized multi-omic profiles and estimate modality-specific cellular compositions, using pseudo-bulk data constructed by internal non-transcriptomic signature matrix recovered from target tissues and external scRNA-seq reference. The accuracy of MODE was evaluated through extensive simulation experiments generating realistic multi-omic data from distinct tissue types. MODE outperformed seven deconvolution pipelines with superior generalizability and enhanced fidelity across five independent datasets, elucidating multi-omic signatures for developmental origins, evolution, subtyping of pediatric medulloblastoma and the prognosis of adult glioblastoma.

Keywords: multimodal, autoencoder, high-resolution purification, origin cell mapping, tumor microenvironment

Title: CoMPaSS: A Computational Pipeline for Cross-Platform Concordance Assessment and Navigating Study Design in Microbiome Research
Author list: Xi Qiao[1,2], Ruitao Liu[3], Daoyu Duan[3], Qian Li[4], Liangliang[3]
Detailed Affiliations:
[1]Department of Internal Medicine, Epidemiology, School of Medicine University of Utah, Salt Lake City, UT, USA; [2]Huntsman Cancer Institute, Salt Lake City, UT, USA; [3]Department of Population and Quantitative Health Sciences, School of Medicine Case Western Reserve University, Cleveland, OH, USA; [4]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA.
Abstract body: Microbiome analysis is essential for understanding microbial interactions and their impact on human health. Advances in next-generation sequencing (NGS) have led to two widely used methods: 16S rRNA gene sequencing and metagenomic shotgun sequencing. While 16S sequencing enables broad taxonomic classification and phylogenetic analysis, shotgun sequencing provides higher taxonomic resolution and functional insights but at a higher cost. However, their comparative efficacy remains uncertain, complicating study design. To address this, we introduce CoMPaSS (Concordance of Microbiome Sequencing Platforms and Study Initiation Strategy), a computational pipeline for navigating microbiome study design. CoMPaSS systematically evaluates sequencing concordance across multiple levels, from community diversity to taxonomic composition, and provides power analysis, sample size estimation, and cost assessment to support study planning. Through extensive simulations and real-world microbiome studies, we found moderate concordance at broader taxonomic levels but significant discrepancies at finer levels and for rare taxa, emphasizing the impact of sequencing method selection on study outcomes. By integrating statistical and computational insights, CoMPaSS helps researchers optimize study design based on scientific and budgetary constraint.
Keywords: 16S, metagenomic shotgun, concordance, power calculation

Title: SEHI-PPI: An End-to-End Sampling-Enhanced Human-Influenza Protein-Protein Interaction Prediction Framework with Double-View Learning
Author list: Qiang Yang[1], Xiao Fan[2], Haiqing Zhao[3], Zhe Ma[4], Megan Stanifer[4], Jiang Bian[5], Marco Salemi[6], and Rui Yin[1, *]
Detailed Affiliations:
[1] Department of Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, FL, USA. [2] Department of Biomedical Engineering, University of Florida, Gainesville, FL, USA. [3]Department of Biochemistry & Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA. [4] Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, USA. [5] School of Medicine, Indiana University, Indianapolis, IN, USA. [6] Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA
Abstract: Influenza continues to pose significant global health threats, hijacking host cellular machinery through protein-protein interactions (PPIs), which are fundamental to viral entry, replication, immune evasion, and transmission. Yet, our understanding of these host-virus PPIs remains incomplete due to the vast diversity of viral proteins, their rapid mutation rates, and the limited availability of experimentally validated interaction data. Additionally, existing computational methods typically struggle with limited high-quality samples and inability for the modeling on the intricate nature of host-virus interactions. To address these challenges, we present SEHI-PPI, an end-to-end framework for human-influenza PPI

prediction. SEHI-PPI integrates a double-view deep learning architecture that captures both global and local sequence features, coupled with a novel adaptive negative sampling strategy to generate reliable and high-quality negative samples. Our method outperforms multiple benchmarks, including state-of-the-art large language models, achieving a superior performance in sensitivity (0.986) and AUROC (0.987). Notably, in a stringent test involving entirely unseen human and influenza protein families, SEHI-PPI maintains strong performance with an AUROC of 0.837. The model also demonstrates high generalizability across other human-virus PPI datasets, with an average sensitivity of 0.929 and AUROC of 0.928. Furthermore, AlphaFold3-guided case studies reveal that viral proteins predicted to target the same human protein cluster together structurally and functionally, underscoring the biological relevance of our predictions. These discoveries demonstrate the reliability of our SEHI-PPI framework in uncovering biologically meaningful host-virus interactions and potential therapeutic targets.
Keywords: Protein-Protein Interaction, Machine Learning, Host-Virus, Double-view Learning

Title: Cyclin D1 induces epigenetic and transcriptional alterations in Multiple Myeloma with t(11;14)(q13;q32)
Author list: Huihuang Yan, PhD[1], Suganti Shivaram, M.B.B.S[2], Hongwei Tang, PhD[2], Hans Anderson[2], Shulan Tian, PhD[1], Michael D Howe[3], Abiola Bolarinwa, MBBS,FMCPath[4], Cinthya Zepeda Mendoza, PhD[5], Stacey Lehman[2], Leif Bergsagel, MD[6], Esteban Braggio, PhD[7], Rafael Fonseca, MD[8], Shaji Kumar, MD[4], Francesco Maura, MD[9], Linda B. Baughn, PhD[2]
Detailed Affiliations
[1]Division of Computational Biology, Mayo Clinic, Rochester, MN 55905, USA; [2]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA; [3]Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA; [4]Division of Hematology, Mayo Clinic, Rochester, MN 55905, USA; [5]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA; [6]Division of Hematology/Oncology, Mayo Clinic, Phoenix, AZ 85054, USA; [7]Division of Hematology and Oncology, Mayo Clinic, Scottsdale, AZ 85259, USA; [8]Division of Hematology, Mayo Clinic, Phoenix, AZ 85054, USA; [9]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.
Abstract: Cyclin D1, encoded by the *CCND1* gene, is primarily known as a major regulator of cell cycle progression. Emerging studies have demonstrated a role for cyclin D1 in regulating gene transcription. In mantle cell lymphoma with t(11;14)(q13;q32) associated with *CCND1::IGH*, cyclin D1 binds strongly to the promoters of highly transcribed genes; its overexpression results in global transcriptional downregulation and activation of a specific gene set. However, the regulatory role for cyclin D1 in t(11;14) multiple myeloma (MM) remains undefined. We hypothesize that cyclin D1 modulates the expression of a distinct gene set through its impact on the landscape of epigenetic modifications. To address this hypothesis, CRISPR/Cas9 was used to generate knockout (KO) line of *CCND1* in U266B1, an MM cell line with a IGH Ea1 super-enhancer inserted centromeric to *CCND1* resulting in *CCND1::IGH*. ChIP-seq (H3K4me3, H3K4me1, and H3K27ac), ATAC-seq, and RNA-seq were performed for KO and wild-type (WT) clones. Differential expression analysis identified 2-fold more genes that were down-regulated than up-regulated in KO compared to WT (467 vs. 214). Pathway analysis revealed an enrichment of TNF-alpha signaling via NF-kappaB, inflammatory response, and apoptosis in association with down-regulated genes, while up-regulated genes are enriched in myogenesis, KRAS signaling, and epithelial mesenchymal transition pathway. In parallel, we identified 2.5% (1,161) H3K27ac peaks and 0.5-1% (359-610) of the peaks from the other three marks that showed significant changes between KO and WT, predominantly in distal regulatory regions. For ATAC- seq and H3K4me1,

about 70% of the differential peaks showed increased signal in KO, in contrast with H3K27ac for which 60% of differential peaks showed decreased signal in KO. The preferential loss of H3K27ac occupancy is consistent with the observation of a higher proportion of genes being down-regulated in KO. These results indicate that knockout of *CCND1* results in a global increase of chromatin accessibility, but a reduction of H3K27ac and gene expression. Further analysis of differential peaks in the promoter regions revealed that, for ATAC-seq and H3K4me3, 47% and 67% of the differential peaks were associated with differentially expressed genes, vs. ~30% for the two enhancer marks. We conclude that knockout of *CCND1* impacts local chromatin state, particularly at enhancer regions, and the transcriptional program. Further studies will identify the genes targeted by the differential enhancers and their possible roles in MM pathogenesis mediated by cyclin D1 overexpression following the t(11;14) event.
Keywords: ChIP-seq; Cyclin D1; Multiple myeloma; RNA-seq; Translocation

Advances in Bioinformatics
August 4th
2:20 PM – 5:20 PM
Room: 301

Chairs: Yu-Chiao Chiu, Juexin Wang

20
Title: Benchmarking Cellular Deconvolution Algorithms to Predict Cell Proportions: A Literature Review
Author list: Ayla Bratton, Ayesha Malik, Jiao Sun, Qian Li and Wei Zhang
Abstract: Computational cellular deconvolution has recently emerged as a powerful alternative to traditional experimental approaches for analyzing RNA-sequencing (RNA-seq) data. Instead of relying solely on physical separation techniques or histological analysis, researchers can now use in silico methods, such as statistical modeling and machine learning, to estimate the cellular composition of complex tissue samples. A key application of these methods is the deconvolution of bulk RNA-seq data using a reference derived from single-cell RNA-seq data. This review focuses on six widely used reference-based deconvolution algorithms: BayesPrism, CIBERSORTx (CSx), DISSECT, Scaden, TAPE, and scpDeconv. Each method employs a different modeling strategy and offers unique strengths, depending on the context of use. To benchmark these algorithms, both pseudobulk RNA-seq data and single-cell reference data were generated from a publicly available human cerebral cortex scRNA-seq dataset. A ground truth cell fraction matrix and a cell type–specific gene expression signature matrix were also constructed from the same dataset to enable evaluation. The pseudobulk samples were input into each deconvolution algorithm, with the corresponding single-cell data used as the reference. After running the models, the estimated cell type proportions were compared against the ground truth to assess performance. The evaluation revealed that while some algorithms excel at accurately estimating cell type proportions, others also provide reliable predictions of gene expression for individual cell types. As computational deconvolution continues to evolve, selecting an appropriate method for a given dataset and biological question remains a critical step in transcriptomic analysis.

Keywords: cellular fraction prediction; cell-type-specific gene expression profile; scRNA-seq

30

Title: Landscape of gene essentiality in cancer cell death pathways
Author list: Shangjia Li, Zhimo Zhu, Chen Yang, Nuo Sun, Lijun Cheng and Lang Li
Abstract: Regulated cell death (RCD), a process that relies on a series of molecular mechanisms, can be targeted to eliminate superfluous, irreversibly damaged, and potentially harmful cells. To better understand how the cell death pathway contributes to cancer therapy, we studied 1150 cancer cells in Dependency Map (DepMap) database for 12 distinct cell death pathways and assessed their gene essentialities. Genes who are essential in 90% or more cancer cell lines are called always essential; or partial essential if falling into (10%, 90%); or rare essential if they are essential in less than 10% of cancer cell lines. Overall, among these 12 cell death pathways, 23, 47, 551 genes were classified as always essential, partial essential, and rare essential, respectively. In two cell death pathways, Parthanatos, and Pyroptosis, all genes were rare essential. Among the other nine cell death pathways, Apoptosis, Autosis, Necroptosis, Efferocytosis, Ferroptosis, Mitotic cell death, Autophagy, Lysosome cell death, MPT driven necrosis and Immunogenic, there are (10, 1, 13, 6, 3, 6, 11, 1,1,0) partial essential genes (PEG), and (2,1,3,1,1,11,4,0,0,1) always essential genes (AEG). As of the date we collected the data, eleven AEGs and eighteen PEGs did not have targted drugs that under-going clinical trials. These cell death pathways essential genes could be viable targets for therapeutic drug development for cancer therapies.
Keywords: cell death pathway; gene essentiality; pan-cancer

Integrative Bioinformatics for Translational and Precision Medicine

August 5th
9:20 AM – 12:20 PM
Room: 350

Chairs: Yuan Liu, Shilin Zhao

26

Title: A novel immune-related risk stratification model to predict prognosis, immunotherapy and chemotherapy response for Neuroblastoma
Author list: Jiamei Xu, Peng Zhao, Jing Qiao, Yan Chang, Xiaoling Mu, Jingxian Wu, Jin Zhu and Xiaohui Zhan
Abstract: Neuroblastoma (NBL) characterized by high morbidity and mortality is a prevalent pediatric cancer originating from neural crest cells. Unsatisfactory prognostic and treatment effects persist due to NBL patients' clinical diversity and individual variations. Despite immunotherapy as a promising therapy has been used in NBL, it still fails in many cases. Thus, there is a strong need to develop an innovative model to optimize therapeutic outcomes and improve patient survival. In this study, the local maximal

quasi-clique merger (lmQCM) algorithm was employed to identify gene co-expression network (GCN) modules, with module 1, linked to immune function, selected for further analysis. A novel immune-related risk stratification model (NIRSM) was developed based on module 1's key genes, demonstrating associations with poor prognosis, immunotherapy and chemotherapy responses, and superior predictive performance compared to age, INSS stage, and MYCN status. The low-risk patients showed enhanced immunotherapy response and higher immune cell infiltration, while module 1 genes exhibited elevated expression in this group. The clinical features like age, INSS stage, and MYCN status differed significantly between two risk groups ($p < 0.001$). Single-cell analysis confirmed the cell-type-specific expression patterns of NIRSM-related genes in immune cells, underscoring the model's biological and clinical relevance. In total, we established a robust model with important implications for predicting NBL prognosis, immunotherapy and chemotherapy response. Our findings not only provide crucial clinical implications for personalized treatment strategies but also offer potential therapeutic targets.
Keywords: NBL; GCN; Immune-related risk stratification model; Prognosis; Immunotherapy; Chemotherapy

1
Title: The Impact of HLA Diversity on Immune Cell Composition, Tumor Mutation Burden, and Cancer Survival
Author list: Judy Bai, Lilly Wei, Kyle Yang, Qing Luo, Yu Liu, Justin Guo, Fenyao Yan, Limin Jiang, Yan Guo and Shilin Zhao
Abstract: HLA molecules play a crucial role in immune responses by influencing antigen presentation and immune cell composition. While previous studies have focused on specific HLA genes or alleles, the impact of overall HLA diversity remains understudied. In this study, we analyzed HLA diversity in relation to immune cell composition, TMB, and mutational signatures across multiple cancers. Higher HLA diversity correlated positively with cytotoxic immune cells, such as CD8+ T cells and activated NK cells ($p = 7.14 \times 10^{-10}$), while correlating negatively with immunosuppressive cells, including monocytes ($p = 6.00 \times 10^{-8}$). HLA diversity generally showed a negative correlation with TMB, except in GBM ($R = 0.15$, $p = 0.02$), where immune suppression may allow highly mutated cells to persist. In LGG ($R = -0.16$, $p = 0.0002$), higher HLA diversity appeared to enhance immune selection against mutated clones. Additionally, HLA diversity was negatively associated with mutational signatures from tobacco (LUSC: $p = 3.07 \times 10^{-5}$, LUAD: $p = 1.18 \times 10^{-5}$), UV exposure (SKCM: $p = 0.04$), and aflatoxin (LIHC: $p = 0.03$), suggesting a role in limiting mutation accumulation. Survival analysis showed that higher HLA diversity improved survival in SKCM (HR: 0.61, $p = 0.0005$) and LUAD (HR: 0.69, $p = 0.02$) but was linked to poorer survival in LGG (HR: 2.09, $p = 0.0001$), likely due to chronic inflammation and immune evasion.
Keywords: HLA; Cancer; Survival; Mutational Signature; Aflatoxin; UV light; Tobacco

11
Title: Horizontal gene transfer networks reveal resistance of plasmid-mediated communication in antibiotic exposure
Author list: Shuai Cheng Li, Lijia Che, Shuai Wang, Yiqi Jiang, Jingwan Wang, Bowen Tan and Xinyao Li
Abstract: Plasmids play an important role in microbial evolution and adaptation, serving as mediators of horizontal gene transfer (HGT) that facilitates the exchange of genetic material across diverse species. We have deduced the first comprehensive plasmid-mediated HGT network using 214,950 plasmid taxonomic units (PTUs) sourced from the IMG/PR database. In this network, taxa serve as vertices, with edges

symbolizing potential gene exchanges facilitated by plasmids. This network demonstrates a hierarchical structure and high robustness. The network edges exhibit strong specificity to particular environments, while they exhibit similarity and generality across various categories of antibiotic-resistance genes (ARGs). Further, we observed a consistent preservation of plasmid-mediated communication ability in gut microbiome after antibiotic exposure in two independent experiments of antibiotic exposure.
Keywords: horizontal gene transfer; antibiotic resistance; metagenomics; network analysis

12
Title: Boolean Network Modeling-Guided Identification of FDA-Approved Drug Combinations for Targeted Treatment Strategies in Head and Neck Cancer
Author list: Pranabesh Bhattacharjee and Aniruddha Datta
Abstract: Head and neck cancer (HNC) presents significant therapeutic challenges due to pathway redundancies and resistance mechanisms. To address this, we developed a Boolean network model integrating key signaling pathways—EGFR, Wnt, Hippo-YAP, MAPK/ERK, and PI3K/mTOR—to systematically assess single and combination drug therapies. Using the Normalized Size Difference (NMSD) metric, we quantified the efficacy of FDA-approved drugs against tumors with multiple mutations. Our simulations identified VT3989 (YAP/TEAD inhibitor) as the most effective monotherapy. Among two-drug combinations, Ulixertinib (ERK inhibitor) and VT3989 exhibited the lowest NMSD, indicating strong synergistic inhibition of MAPK and Hippo pathways. Adding Vorinostat (FBXW7 modulator) further enhanced efficacy, achieving 80% efficacy. The most effective combination—Temsirolimus (mTOR inhibitor), Ulixertinib, VT3989, and Vorinostat—demonstrated an 88.3% improvement over untreated conditions. Our findings support a shift from sequential to concurrent multi-pathway targeting, mirroring clinical evidence that combination approaches delay resistance. The hierarchical NMSD reductions from 0.685 (single-agent) to 0.120 (four-drug therapy) highlight the advantage of combination depth in pathway control. This computational framework provides a rationale for prioritizing Temsirolimus-containing quadruple therapies, offering a novel precision oncology strategy for HNC with complex mutational landscapes.
Keywords: Boolean Network; Combination Therapy; Drug Repurposing; Head and Neck Cancer; Targeted Therapy

27
Title: Comparison of Nanopore Sequencing, MethylationEPIC Array, and EM-Seq for DNA Methylation Detection
Author list: Steven Brooks, Hongyu Gao, Xuhong Yu, Yunlong Liu and Gang Peng
Abstract: DNA Methylation is an important biological process in epigenetics, and many methods have been developed to profile DNA methylation. An increasing number of recent studies have employed Oxford Nanopore long-read sequencing technology for DNA methylation detection, presenting an alternative to the widely utilized Infinium arrays and short-read whole-genome sequencing methods. In this study, we evaluate the performance of Nanopore sequencing in DNA methylation detection by comparing it to the Illumina's MethylationEPIC microarray (EPIC) and Enzymatic Methyl-Sequencing (EM-Seq). The initial comparison was conducted between the Nanopore platform and the EPIC array. Among the ~850,000 CpG sites covered by both methods, we observed high concordance (Pearson correlation coefficient, $r \geq 0.94$ across all four samples). After downsampling Nanopore data from an average coverage of 25.9 reads per site to 10 reads per site, the correlation in CpG methylation remained high ($r \geq 0.93$). Lower correlation of CpG methylation (r: 0.79 - 0.88) was detected between Nanopore and EM- Seq, which can be attributed to

biased and reduced coverage of hypomethylated CpG sites by EM-Seq. We also investigated new features detected by Nanopore sequencing, such as native DNA sequencing that can differentiate 5mC and 5hmC, as well as haplotype-specific methylation. Overall, the Nanopore platform exhibited a high degree of concordance with the EPIC array and provided more uniform genomic coverage than EM-Seq. This study provides insights for researchers in selecting appropriate DNA methylation detection methods, considering factors such as cost, DNA input, and the complexity of downstream analysis.
Keywords: Nanopore; 5mc/5hmc; haplotype-specific methylation

51
Title: A Hierarchical Adaptive Diffusion Model for Flexible Protein-Protein Docking
Author list: Rujie Yin and Yang Shen
Abstract: Protein–protein interaction prediction is critical but challenged by significant conformational changes. We propose a hierarchical adaptive diffusion model separating global rigid-body motions and local flexibility, with noise schedules mimicking induced fit effects. Local flexibility is adaptively conditioned on predicted conformational change levels. Using the DIPS-AF dataset, the model outperformed AlphaFold2-like and DiffDock-PP models, especially in flexible cases, demonstrating improved docking accuracy.
Keywords: Protein docking; Conformational changes; Generative models; Diffusion models

25
Title: Knowledge-driven annotation for gene interaction enrichment analysis
Author list: Xiaoyu Liu, Anna Jiang, Chengshang Lyu and Lingxi Chen
Abstract: Gene Set Enrichment Analysis (GSEA) is a cornerstone for interpreting gene expression data, yet traditional approaches overlook gene interactions by focusing solely on individual genes, limiting their ability to detect subtle or complex pathway signals. To overcome this, we present GREA (Gene Interaction Enrichment Analysis), a novel framework that incorporates gene interaction data into enrichment analysis. GREA replaces the binary gene hit indicator with an interaction overlap ratio, capturing the degree of overlap between gene sets and gene interactions to enhance sensitivity and biological interpretability. It supports three enrichment metrics: Enrichment Score (ES), Enrichment Score Difference (ESD) from a Kolmogorov-Smirnov-based statistic, and Area Under the Curve (AUC) from a recovery curve. GREA evaluates statistical significance using both permutation testing and gamma distribution modeling. Benchmarking on transcriptomic datasets related to respiratory viral infections shows that GREA consistently outperforms existing tools such as blitzGSEA and GSEApy, identifying more relevant pathways with greater stability and reproducibility. By integrating gene interactions into pathway analysis, GREA offers a powerful and flexible tool for uncovering biologically meaningful insights in complex datasets. The source code is available at https://github.com/compbioclub/GREA.
Keywords: Bioinformatics; Gene Set Enrichment Analysis; Transcriptome

22
Title: “Frustratingly easy" domain adaptation for cross-species transcription factor binding prediction
Author list: Mark Maher Ebeid, Ali Tugrul Balcı, Maria Chikina, Panayiotis V Benos and Dennis Kostka
Abstract: Motivation: Sequence-to-function models, designed to interpret genomic DNA and predict functional outputs, have demonstrated success in characterizing regulatory sequence activity. However, interpreting these models remains an open challenge, raising questions about whether they learn

generalizable biochemical properties. Cross-species prediction of transcription factor (TF) binding offers a promising avenue to push models toward generalization, leveraging variation across species to potentially uncover a conserved regulatory code. Nonetheless, accounting for systematic differences between the genomes of different species presents a significant challenge.
Results: We introduce MORALE, a framework leveraging an established domain adaptation approach that is “frustratingly easy". MORALE trains on sequences from one or more source species and predicts TF binding on a single target species; in order to learn an invariant cross-species representation, MORALE simultaneously aligns the moments (i.e., 1st and 2nd) between all species. This direct approach integrates readily into models with an embedding layer. Unlike adversarial alternatives, it requires no additional parameters and does not alter the standard gradient computation. We apply MORALE to two ChIP-seq datasets of liver-essential TFs: one comprising human and mouse, and another comprising five mammalian species. Compared to both the baseline and gradient reversal (GRL), MORALE demonstrates improved performance across all TFs in the two-species case, avoiding the performance degradation observed with the GRL approach in this study. Furthermore, gradient inspection revealed that the de novo motifs discovered by MORALE adhered more strictly to CTCF compared to the GRL approach. For the five-species case, our method significantly improved TF binding site prediction for all TFs when predicting on human data, surpassing the performance of a human-only model — a result not observed in the two-species comparison. Overall, MORALE is a direct and competitive approach that leverages domain adaptation techniques to improve cross-species TF binding site prediction.
Keywords: unsupervised domain adaptation; regulatory genomics; transcription factor binding site prediction; moment alignment; invariant representation learning

AI and Machine Learning in Translational Genomics

August 5th
1:30 AM – 4:50 PM
Room: 350

Chairs: Huihuang Yan, Yixing Han

53
Title: Adaptive Chebyshev Graph Neural Network for Cancer Gene Prediction with Multi-Omics Integration
Author list: Li Sa
Abstract: Identifying cancer driver genes is computationally challenging due to diverse genetic and non-genetic factors. We present ACGNN, integrating pan-cancer multi-omics data and PPI networks into graph convolutional networks refined with adaptive Chebyshev filters for flexible feature aggregation. ACGNN achieved a 25.9% AUPRC improvement over state-of-the-art methods, accurately identifying

established and novel cancer driver genes, providing valuable insights for cancer research and precision medicine.
Keywords: Cancer driver genes; Graph neural network; Node embeddings; Chebyshev networks

54
Title: In Silico Design of a Population-Specific mRNA Vaccine Targeting MUC1 for Colorectal Cancer: Focus on Iranian HLA Diversity
Author list: Sara Farahbakhsh, Zarrin Minuchehr, Raha Mahdavi Karimi, Hanita Kouzegar, Atrin Tofighi, and Amitis Masumian
Abstract: Advances in mRNA vaccines offer new cancer immunotherapy strategies. We designed an Iranian population-specific mRNA vaccine targeting MUC1, overexpressed in colorectal cancer. Bioinformatics identified common HLA alleles (HLA-A*24:02, HLA-B*35:01, HLA-C*04:01) and immunodominant epitope SVSDVPFPF with strong binding and high immunogenicity. The optimized vaccine demonstrated high stability and 98.3% population coverage, highlighting its potential as a targeted immunotherapy for MUC1-associated cancers and a framework for developing similar vaccines.
Keywords: mRNA vaccine; MUC1; Colorectal cancer; Bioinformatics

2
Title: A Generative Imputation Method for Multimodal Alzheimer's Disease Diagnosis
Author list: Reihaneh Hassanzadeh, Anees Abrol, Hamid Reza Hassanzadeh and Vince D. Calhoun
Abstract: Multimodal data analysis can lead to more accurate diagnoses of brain disorders due to the complementary information that each modality adds. However, a major challenge of using multimodal datasets in the neuroimaging field is incomplete data, where some of the modalities are missing for certain subjects. Hence, effective strategies are needed for completing the data. Traditional methods, such as subsampling or zero-filling, may reduce the accuracy of predictions or introduce unintended biases. In contrast, advanced methods such as generative models have emerged as promising solutions without these limitations. In this study, we proposed a generative adversarial network method designed to reconstruct missing modalities from existing ones while preserving the disease patterns. We used T1-weighted structural magnetic resonance imaging and functional network connectivity as two modalities. Our findings showed a 9% improvement in the classification accuracy for Alzheimer’s disease versus cognitive normal groups when using our generative imputation method compared to the traditional approaches.
Keywords: Generative Adversarial Networks; Multi-Modal Classification; Alzheimer's Disease

13
Title: A user-friendly R Shiny app for Predicting Surface Protein Abundance from scRNA-seq Expression Using Deep Learning in blood cells
Author list: Hui-Mei Tsai, Tzu-Hung Hsaio, Yu-Chiao Chiao, Eric Y. Chuang and Yidong Chen
Abstract: Understanding accurate immune cell heterogeneity and function in single-cell datasets requires access to protein-level information, which is often missing due to experimental limitations. To address this gap, we present shinyDeepGxP, an interactive web application that implements our deep learning model, DeepGxP, for predicting surface protein abundance from single-cell RNA-sequencing (scRNA-seq) data. The platform makes DeepGxP accessible to researchers without programming expertise. Users can upload scRNA-seq count matrices and perform “Predict Protein”, which predicts the abundance of 224 biologically relevant surface proteins. shinyDeepGxP also provides visualization to aid in identifying distinct cell populations based on predicted protein profiles. Moreover, users can access to “Interpret Model”, which

reveals key RNA predictors and their associated biological pathways for each protein. In summary, shinyDeepGxP is a user-friendly and freely available web tool that brings protein-level resolution to RNA-only single-cell datasets, supporting multimodal discovery without the need for additional experiments.
Keywords: Shiny web; Protein prediction; Single-cell; Deep learning

44
Title: HELP-TCR Harmonized Explainable Language Processing Toolkit for T-Cell Antigen Receptor Repertoires
Author list: Michal Seweryn, Yulyana Kalesnik, Dawid Krawczyk, and Maciej Pietrzak
Abstract: Functional characterization of T-cell antigen receptor (TCR) repertoires is critical for advancing our understanding of adaptive immune responses across diverse contexts, including infectious diseases, cancer, autoimmune conditions, and allergic disorders. Detailed analysis of TCR repertoires can reveal disease-specific signatures, support biomarker discovery, and facilitate the development of immunotherapies and vaccines. However, current computational approaches often prioritize either global repertoire metrics or employ deep learning models that, while powerful, offer limited interpretability. We introduce HELP-TCR, a novel machine learning framework based on natural language processing that combines low-dimensional, explainable feature extraction with robust classification performance. HELP-TCR represents TCR repertoires by modeling the position-specific distributions of single amino acids and amino acid pairs, transforming sequences into multidimensional tensor structures. To increase reproducibility, a consensus grouping method merges features with highly similar position-wise distributions. A modified ResNet-18 deep learning architecture, adapted to process these tensors, enables accurate classification, while post-hoc saliency map analysis highlights the most informative features contributing to model predictions. Using a dataset of bootstrapped TCR sequences, HELP-TCR achieved an AUC of 0.96, outperforming existing methods including DeepTCR (AUC 0.76) and TCR-BERT embeddings. Beyond performance, HELP-TCR enables identification of position-specific amino acid motifs associated with classification decisions, offering biologically interpretable insights into TCR repertoire differences. By emphasizing model interpretability alongside predictive accuracy, HELP-TCR provides a versatile platform for functional TCR repertoire analysis with potential applications in immunotherapy development, vaccine design, and immune monitoring.
Keywords: T-cell antigen receptors; explainable AI; deep learning; neural network-based classification; Wasserstein distance

23
Title: Efficient and Valid Large Molecule Generation via Self-supervised Generative Models
Author list: Doyoung Kwak, Raiyan Chowdhury, Byung-Jun Yoon and Xiaoning Qian
Abstract: The realm of molecular design, particularly for large molecules, presents unique challenges and opportunities in drug discovery and materials science. Large molecule design is inherently more complex and less explored compared to designing small molecules, adding significant difficulty in generative modeling. We aim to establish strong baselines for better scalability, efficiency, and generative performance in this domain. We evaluate the scalability and performance of generative AI models, initially effective for small molecule design, in generating large molecules for potential drugs in gene-based therapies, immunotherapies, hormonal regulators, and targeted cancer therapies. Our findings indicate that computational strategies and model architectures designed for small molecules may not readily extend to large molecular structures. To address these limitations, we explore masked language modeling strategies alongside advanced tokenization methods, including Atom-Pair Encoding (APE), to enhance generative AI

models. We probe how incorporating such strategies, particularly the APE tokenization method that explicitly captures structural and chemical characteristics, can significantly improve design capabilities for complex molecular structures. Overall, our results demonstrate both the potential and challenges of deep generative modeling for large molecules and how the proposed enhancements may bridge the gap in generating large molecules when novel discovery is the ultimate goal.
Keywords: Molecule generation; Large molecule; Generative model; Molecule representation; Self-supervised Learning; String representation; SMILES; SELFIES; Tokenization; BPE; APE

47
Title: DG-scRNA: Deep Learning with Graphic Cluster Visualization to Predict Cell Types of Single-Cell RNAseq Data
Author list: Birkan Gokbag, Yimin Liu, Abhishek Majumdar, Lang Li, Chongwen Dong, Yanan Song, Wei Xia, and Lijun Cheng
Abstract: Single-cell RNA sequencing (scRNA-seq) has revolutionized understanding of cellular heterogeneity, but accurate cell type annotation remains challenging. Marker genes, essential for distinguishing cell types, vary by tissue origin, disease state, and experimental methods. Here, we present DG-scRNA, a deep learning framework with graphic cluster visualization that systematically optimizes marker selection based on sample context, improving cell type identification. Validated with T-cell annotation from thyroid cancer scRNA-seq data, DG-scRNA achieved superior performance (F1 score: 95.19%) compared to existing annotation methods. Its automated, context-specific marker selection matches optimal markers based on species, tissue, and disease state. Applying DG-scRNA to papillary thyroid carcinoma samples uncovered distinct T-cell subpopulations associated with metastasis patterns. DG-scRNA offers an accurate, context-aware solution for cell type annotation across diverse systems and disease settings.
Keywords: Thyroid cancer; Single-cell RNA sequencing (scRNA-seq); Cell type annotation

50
Title: A Machine Learning–Enhanced Pipeline for Detecting Disruption of Transcription Termination (DoTT) in RNA-Seq Data
Author list: Michael Levin, Igor Astsaturov, and Yunyun Zhou
Abstract: Disruption of transcription termination (DoTT) occurs when RNA polymerase II fails to stop at the 3′ end of a gene, producing readthrough transcripts often missed by standard RNA-seq pipelines. We developed a pipeline that extends gene annotations downstream, quantifies readthrough reads, and applies differential expression analysis. A Random Forest classifier further boosts sensitivity (25% → 65%). Applied to high-carbohydrate diet and HSV-1 infection datasets, the pipeline detected ~50 and 707 DoTT events, respectively, recovering ~78% of known HSV-1 events. Pathway analysis revealed immune-related pathways, highlighting the pipeline's broad applicability and improved DoTT detection.
Keywords: Transcriptional readthrough; Transcription termination failure; RNA-seq; High-carbohydrate diet; HSV-1; Machine learning; Random Forest; DoGFinder; Readon

Title: THANOS: An AI Pipeline for Engineering Antibodies
Author list: Arnav Solanki1, Neha S Maurya1, Wenjin Jim Zheng1
Detailed Affiliations:
1McWilliams School of Bioinformatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract: The Brain Research Through Advancing Innovative Neurotechnologies (BRAIN) initiative seeks to unravel the complexities of brain cell types, connections, and functions. One powerful approach to studying neural circuits is imaging brain proteins using antibodies. However, generating high-quality antibodies is often slow and expensive. Recent advances in AI tools like AlphaFold3 and RFdiffusion offer a fast, fully digital alternative to traditional experimental screening. This project introduces THANOS (Targeted High-throughput Antibody Notation, Optimization, & Screening), a novel pipeline for rapidly engineering. THANOS was used to design de novo antibodies targeting human proteins by redesigning antigen-binding sites of antibodies using a high-performance GPU server.
This case study focuses on Parvalbumin (PVALB), a calcium-binding protein abundant in neural cells. Alphafold3 was used to model the 3D complexes of human PVALB and 12 mouse antibody variable fragments. With these 12 models as initial states, 300 structural variants were generated by using RFdiffusion to diffuse the complementarity determining regions (CDRs) at the binding sites. ProteinMPNN was used to predict optimal sequences that fold these structures, yielding new antibody chain sequences containing new residues in the CDRs. These 300 variants were screened against PVALB using AlphaFold3 to predict their binding. 4 candidates were observed to bind strongly based on low predicted alignment error. These candidates were validated through:

1) Structural inspection using ChimeraX.
2) Molecular dynamics simulations with GROMACS. The best candidate demonstrated a MMGBSA energy of -35 kcal/mol over 100 ns.
3) Solubility checks using Aggrescan3D to confirm the absence of aggregation-prone residues.
4) Sequence alignment to ensure minimal mutations and preserve nativeness.

These antibodies are currently undergoing experimental validation. THANOS demonstrates how AI can accelerate antibody engineering from weeks to hours without a wet lab. This pipeline can be applied on any desired target protein (beyond the interest of the BRAIN initiative) for numerous applications such as viral or cancer therapy, and will be invaluable to the fields of immunology and pharmacology.
Keywords: Antibodies, Protein Engineering, AI, Machine Learning, AlphaFold

7

Title: DisSubFormer: A Subgraph Transformer Model for Disease Subgraph Representation and Comorbidity Prediction

Author list: Ashwag Altayyar and Li Liao

Abstract: Considering the complexity of diseases, comorbidity arises from intricate molecular interactions, functional relations, and shared pathological mechanisms, making comorbidity prediction a challenging yet prominent research topic in bioinformatics. As disease etiologies are inherently multifaceted, integrating multi-source data, such as the protein-protein interaction (PPI) network and Gene Ontology (GO), is crucial for identifying potential comorbid diseases. In this work, we develop DisSubFormer, a novel framework that combines GO-derived semantic features with PPI-based molecular interactions to generate biologically enriched representations of disease-associated proteins. These representations are leveraged within a subgraph Transformer model to represent disease subgraphs by capturing both local structural patterns and global relational information within the PPI network. More specifically, to enhance the scalability of the subgraph Transformer model on a large-scale PPI network, we introduce a biologically informed anchor patch sampling strategy integrated with a head-specific relational attention mechanism to learn context-aware disease subgraph representations for comorbidity prediction while simultaneously reducing computational complexity. We evaluate our proposed method on a benchmark dataset, achieving superior

performance compared to state-of-the-art methods in disease comorbidity prediction, with an AUROC of 0.97.
Keywords: diseases; comorbidity; subgraph Transformer; protein-protein interaction; gene ontology; subgraph embedding; Hawkes process

32
Title: GRN-Integrated Heterogeneous Attentive Graph Autoencoder for Cell-Cell Interaction Reconstruction from Spatial Transcriptomics
Author list: Aiwei Yang, Yujian Lee, Yue Guan and Jiaxing Chen
Abstract: The reconstruction of cell-cell interaction networks (CCIs)
is pivotal to unraveling the regulatory mechanisms governing orchestrated multicellular systems. While deep learning models show advances in inferring CCIs from spatial transcriptomes; most current approaches neglect intracellular gene regulatory dynamics that mediate communication or only get to suboptimal integration of spatial cellular contexts with gene regulatory networks (GRNs), while neglecting the potential data noise amplification through model process. To address these challenges, we propose SFNET, a robust graph neural framework that synergizes noise-resilient graph construction with topology-aware deep learning.
Our Shared Factor Neighborhood (SFN) algorithm constructs cell graphs through joint optimization of spatial coordinates and genetic features, better reducing noise sensitivity compared to conventional KNN approaches.The Heterogeneous Attention Embedding (HAE) module then explicitly models cell-GRN interactions via multi-head cross-domain attention to preserve biological specificity.
Finally, our Triple-Enhancement Graph Neural Network (CECB/SEB/EEB) combats feature degradation through multi-scale enhancement blocks. Enabling precise modeling of both local and long-range interactions in heterogeneous networks. When benchmarked versus existing models, SFNET reduces training time by 50% without compromising accuracy. It improves the average precision (AP) by 2.02%, ROC by 2.56%, and AUROC by 1.72%, highlighting the gains in both accuracy and computational efficiency in CCI inference.
Keywords:
Cell-cell Interaction Network; Graph Neural Network; Shared Neighbourhood Algorithm; Single-cell Spatial Transcriptomics

Data-Driven Insights into Disease Modeling

August 5th
1:30 AM – 4:50 PM
Room: 350

Chairs: Shulan Tian, Joseph McElroy

16

Title: Latent factor modeling reveals unexpected spatial heterogeneity in human Alzheimer's disease brain transcriptomes
Author list: Rami Al-Ouran, Chaozhong Liu, Linhua Wang, Ying-Wooi Wan, Chaohao Gu, Xiqi Li, Gerarda Cappuccio, Mirjana Maletic-Savatic, Aleksandar Milosavljevic, Joshua Shulman, Hu Chen and Zhandong Liu
Abstract: Alzheimer's disease is characterized by complex molecular and cellular heterogeneity, which complicates efforts to identify consistent biomarkers and therapeutic targets. To gain a deeper understanding of the heterogeneity, we applied latent factor modeling to RNA-seq data from approximately 2,500 human Alzheimer's disease brain samples, uncovering underlying patterns in gene expression. These transcriptional groups demonstrated unique gene expression profiles related to synaptic and neuronal pathways, vasculature development, and protein folding and antigen processing. We demonstrated that this latent factor emerges from variations in spatial sampling. Adjusting for the latent factor recovers nearly three times more differentially expressed genes than analyses not stratified by this factor. This finding suggests that spatial heterogeneity is a pervasive element across various cellular and molecular brain profiles and has far-reaching implications for future studies of Alzheimer's disease and related neurological disorders.
Keywords: Brain transcriptome; Latent factor; Sampling variations

48
Title: Compositional Bayesian Co-Clustering of DTI Biomarkers with Clinical Measures for Enhanced Prediction of Parkinson Disease Severity
Author list: Ashwin Vinod and Chandrajit Bajaj
Abstract: Parkinson's disease (PD) exhibits inter-patient heterogeneity complicating prognosis and precision therapy. We propose an end-to-end Compositional Bayesian Co-Clustering (SRVCC) framework that integrates tissue-specific diffusion-tensor imaging (DTI) biomarkers and clinical assessments to uncover multimodal patterns predictive of disease severity. Four-dimensional DTI scans from the Parkinson's Progression Markers Initiative were processed to generate extracellular-contamination-free fractional anisotropy and mean diffusivity. Combined with clinical scores (UPDRS and MoCA), SRVCC jointly clusters imaging and clinical data in a variational latent space, suppressing noise while preserving discriminative latent modes. Cross-validation identified three patient subtypes corresponding to mild, intermediate, and severe PD, with imaging metrics mirroring progression. SRVCC outperformed k-means, spectral bi-clustering, and deep-clustering baselines, offering biologically consistent clusters differentiating cognitive status, motor severity, and microstructural changes, bridging DTI alterations and clinical manifestations.
Keywords: DTI imaging; Parkinson's Disease; Compositional Bayesian Co-Clustering

31
Title: DuAL-Net: A Hybrid Framework for Alzheimer's Disease Prediction from Whole Genome Sequencing via Local SNP Windows and Global Annotations
Author list: Eun Hye Lee, Taeho Jo
Abstract: Alzheimer's disease (AD) dementia is the most common formof dementia. With the emergence of disease-modifyingtherapies for AD such as anti-amyloid monoclonalantibodies, the ability to predict disease risk beforesymptom onset has become increasingly important. Wholegenome sequencing (WGS) data is a promising data form forearly AD prediction, despite several analytical challenges.In this study, we introduce DuAL Net, a hybrid deeplearning framework designed to predict

AD dementia usingWGS data.DuAL-Net integrates two components: local probabilitymodeling, which segments the genome into non-overlappingwindows, and global annotation-based modeling, whichannotates each SNP and reorganizes the WGS input to capturelong range functional relationships. Both components employout of fold stacking with TabNet and Random Forestclassifiers. The final prediction is generated by combininglocal and global probabilities using an optimized weightingparameter alpha.We applied DuAL-Net to WGS data from 1,050 individuals (443cognitively normal and 607 with AD dementia), usingfive-fold cross validation for training and evaluation. Onaverage across the 100, 500, and 1000 SNP subset sizesevaluated, DuAL-Net achieved an Area Under the Curve (AUC)of 0.671 using top-ranked SNPs prioritized by the model,representing 35.0% and 20.3% higher predictive performancecompared to the average AUCs of bottom-ranked and randomlyselected SNPs, respectively. Assessment of modeldiscriminative ability via ROC analysis across differentSNP subset sizes consistently demonstrated a strongpositive correlation between the SNPs' prioritization rankand their predictive power. The model identified SNPs withknown associations to AD as top contributors to prediction,alongside potentially novel variants also ranked highly bythe model.In conclusion, DuAL Net presented a promising framework forAD prediction that improved predictive accuracy andenhanced biological interpretability. The framework and itsweb-based implementation offer an accessible platform forbroader research applications.

Keywords: Alzheimer's disease; Disease prediction; Whole genome sequencing; Deep learning; Machine learning

14

Title: Resolving Gene Heterogeneity in DEG Analysis: A Novel Pipeline for Precision Genomics

Author list: Jiasheng Wang, Ilia Buralkin, Rami Ai-Ouran and Zhandong Liu

Abstract: Gene heterogeneity, driven by extensive genetic variation across samples, poses significant challenges in identifying disease-associated genes through Differentially Expressed Gene (DEG) analysis. Traditional global DEG methods often fail to capture subtle yet biologically meaningful signals, which are obscured by genetic variability. To address this, we developed a novel DEG analysis pipeline that integrates DNA and RNA data to amplify signals within genetically similar local regions. This approach combines dimensionality reduction techniques, local contrast identification, and co-regulation modeling to uncover subgroup-specific DEG signals. Our results demonstrate that this local analysis method significantly outperforms traditional global DEG approaches, particularly in detecting weak signals or those localized to small subgroups of samples. Applying this pipeline to the ROSMAP dataset identified key Alzheimer's disease (AD)-related pathways, such as ATP biosynthesis, nervous system development, and synaptic signaling, which were missed by conventional methods. Cross-dataset validation further confirmed the robustness of our approach, showing improved consistency in DEG detection and capturing a broader spectrum of gene expression changes. This study highlights the importance of addressing genomic heterogeneity in DEG analysis and offers a powerful tool for uncovering biologically relevant pathways and disease mechanisms. The proposed method has broad implications for precision medicine, enabling the identification of subgroup-specific signals in complex, heterogeneous datasets.

Keywords: Gene Heterogeneity; Alzheimer's Disease; Differentially Expressed Genes

29

Title: Multimodal Imaging and Cell-Free DNA Methylation Analysis for Noninvasive Lung Cancer Diagnosis

Author list: Ran Hu, Stephen Park, Paul Li, Weihua Zeng, Yonggang Zhou, Chun-Chi Liu, Shuo Li, Xiaohui Ni, Kostyantyn Krysan, Steven Dubinett, Denise Aberle, Ashley Prosper, Wenyuan Li, William Hsu and Xianghong Zhou

Abstract: Background: Low-dose computed tomography (LDCT) is an effective noninvasive screening tool for lung cancer. However, imaging-detected lesions often require invasive follow-up procedures for definitive diagnosis, increasing healthcare costs and the risk of overdiagnosis. There is a pressing need for additional noninvasive methods to improve diagnostic accuracy in patients with CT-detected lung lesions. Objectives: This study aims to develop a multimodal approach that integrates CT imaging and cell-free DNA (cfDNA) methylome data for noninvasive lung cancer diagnosis.

Methods: We utilized large single-modality datasets to pretrain deep learning (DL) models that extract robust and informative features from high-dimensional imaging and methylome data. For imaging, a foundation model was fine-tuned on 677 lung CT lesions to improve its ability to capture lung-specific imaging patterns. Handcrafted radiomic features were also extracted from the CT scans. For cfDNA methylation, lung cancer-specific biomarkers were first identified, followed by training an autoencoder model on 513 normal and lung cancer plasma samples to generate meaningful methylation feature embeddings. After feature extraction, we integrated the multimodal features using early, intermediate, and late fusion strategies, and evaluated model performance on 77 individuals with paired imaging and methylome data using support vector machine (SVM) and neural network (NN) classifiers under 5-fold cross-validation (CV).

Results: On this multimodal dataset, intermediate fusion of imaging and methylation features achieved the highest area under the receiver operating characteristic curve (AUC = 0.870 ± 0.128). The model also demonstrated strong performance in early-stage lung cancer detection, achieving an AUC of 0.806 for Stage I cases versus controls with CT-detected benign lesions.

Conclusions: Integrating multimodal imaging and cfDNA methylation features enhances the accuracy of lung cancer diagnosis and holds promise as a noninvasive approach for distinguishing malignant from benign CT-detected lesions.

Keywords: lung cancer; foundation model; deep learning; machine learning; medical imaging; cell-free DNA; methylation; multimodal data

39

Title: Multidimensional Impact of Microbiota Absence on Thymic T Cell Development in Mice: A Study Based on Single-Cell and Spatial Transcriptomics

Author list: Yifei Sheng, Qian Zhang, Zhao Zhang and Juan Shen

Abstract: Background: Gut microbiota plays an important role in host immune development, but the mechanisms of its influence on primary lymphoid organs such as the thymus remain unclear. Germ-free mice provide an ideal model for studying the impact of microbiota absence on thymus development.

Methods: This study utilized single-cell transcriptomics and spatial transcriptomics techniques to systematically compare thymic cellular composition and gene expression characteristics between germ-free mice and specific pathogen-free mice at different developmental stages (0, 2, 4, and 10 weeks of age).

Results: Spatial transcriptome analysis revealed that GF mouse thymus had 5 basic transcriptome classifications and lacked plasma cells and neutrophils, while SPF mouse thymus had 7 basic classifications. Although early T cell development was not affected under germ-free conditions (no significant differences in DN T cell numbers and expression of activation marker Itgal), T cell

proliferation in pre-pubescent GF mice was significantly lower than in SPF mice, a difference that largely disappeared after puberty (10 weeks). Longitudinal analysis found that with increasing age, double-positive T cells gradually decreased while immature T cells increased. GF mice exhibited more pronounced Th1/Th2 imbalance and greater cell number fluctuations compared to SPF mice. Additionally, PLZF+ innate lymphoid cells in GF mice showed impaired early development but significant increase at 10 weeks of age, possibly representing a compensatory mechanism. Aire expression in thymic mesenchymal cells was significantly lower in GF mice compared to SPF mice. Conclusion: The impact of microbiota absence on thymic T cell development exhibits time- and cell type-specific patterns. While early T cell development remains unaffected, microbiota absence leads to restricted T cell proliferation, Th1/Th2 imbalance, abnormal innate lymphoid cell development, and decreased Aire expression. These findings provide new perspectives for understanding the role of microbiota in shaping the host immune system.
Keywords: Germ-free mice; thymus;T cell development; single-cell transcriptomics;spatial transcriptomics

Title: The Drug Overdose Surveillance in Ohio: What we can see with the geospatial shared component analysis of the Urine Drug Test Results.

Author list: Joanne Kim[1], John Myers[1], Charles Marks[2], Penn Whitley[2], Brandon Slover[1], Xianhui Chen [1], Neena Thomas[1], Ping Zhang[1], Naleef Fareed [1], Soledad Fernandez[1].
Detailed Affiliations: [1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC;
Abstract: Geospatial analysis of the substance use disorder (SUD) population has provided various insights for the surveillance of the SUD population. Numerous data sources have been investigated but the chronic challenge regarding delayed reporting and the scarcity of the data still remains.
To overcome this challenge, we conducted the Bayesian multivariate spatiotemporal modeling analysis using the real-time Urine drug test results for diverse sets of drugs (e.g. Fentanyl, Cocaine, Heroine and Methamphetamine). We use the multivariate Bayesian spatiotemporal approach to investigate the shared geospatial pattern of the substance use population. By looking at their shared components, we can investigate the co-evolving pattern of the drug substance use population in each county from 2013 to 2023. With this effort, we can confirm the existing belief about polysubstance use, and identify new shared patterns with newly emerged substances. We also expect information sharing of multiple drugs can help improve the estimation results of small areas. This talk will discuss the analysis results for various sets of drugs and how the map of substance use population changes in the 10-year period in Ohio.
Keywords: Opioid overdose, Bayesian methods, shared component model, Urine Drug Test, Spatiotemporal

19

Title: AutoRADP: An Interpretable Deep Learning Framework to Predict Rapid Progression for Alzheimer's Disease and Related Dementias Using Electronic Health Records

Author list: Qiang Yang, Weimin Meng, Pei Zhuang, Stephen Anton, Yonghui Wu and Rui Yin
Abstract: Alzheimer's disease (AD) and AD-related dementias (ADRD) exhibit heterogeneous progression rates, with rapid progression (RP) posing significant challenges for timely intervention and treatment. The increasingly available patient-centered electronic health records (EHRs) have made it possible to develop advanced machine learning models for risk prediction of disease progression by leveraging comprehensive clinical, demographic, and laboratory data. In this study, we propose AutoRADP,

an interpretable autoencoder-based framework that predicts rapid AD/ADRD progression using both structured and unstructured EHR data from UFHealth. AutoRADP incorporates a rule-based natural language processing method to extract critical cognitive assessments from clinical notes, combined with feature selection techniques to identify essential structured EHR features. To address the data imbalance issue, we implement a hybrid sampling strategy that combines similarity-based and clustering-based upsampling. Additionally, by utilizing SHapley Additive exPlanations (SHAP) values, we provide interpretable predictions, shedding light on the key factors driving the rapid progression of AD/ADRD. We demonstrate that AutoRADP outperforms existing methods, highlighting the potential of our framework to advance precision medicine by enabling accurate and interpretable predictions of rapid AD/ADRD progression, and thereby supporting improved clinical decision-making and personalized interventions.
Keywords: Alzheimer's Disease and Related Dementias; Deep learning; Interpretable learning; Electronic Health Records; Data Imbalance

33
Title: Machine Learning-Based Mortality Prediction in Critically Ill Patients with Hypertension: Comparative Analysis, Fairness, and Interpretability
Author list: Shenghan Zhang, Sirui Ding, Zidu Xu and Jiancheng Ye
Abstract:
Background: Hypertension is a leading global health concern, significantly contributing to cardiovascular, cerebrovascular, and renal diseases. In critically ill patients, hypertension poses increased risks of complications and mortality. Early and accurate mortality prediction in this population is essential for timely intervention and improved outcomes. Machine learning (ML) and deep learning (DL) approaches offer promising solutions by leveraging high-dimensional electronic health record (EHR) data.
Objective: To develop and evaluate ML and DL models for predicting in-hospital mortality in hypertensive patients using the MIMIC-IV critical care dataset, and to assess the fairness and interpretability of the models. Methods: We developed four ML models—gradient boosting machine (GBM), logistic regression, support vector machine (SVM), and random forest—and two DL models—multilayer perceptron (MLP) and long short-term memory (LSTM). A comprehensive set of features, including demographics, lab values, vital signs, comorbidities, and ICU-specific variables, were extracted or engineered. Models were trained using 5-fold cross-validation and evaluated on a separate test set. Feature importance was analyzed using SHapley Additive exPlanations (SHAP) values, and fairness was assessed using demographic parity difference (DPD) and equalized odds difference (EOD), with and without the application of debiasing techniques.
Results: The GBM model outperformed all other models, with an AUC-ROC score of 96.3%, accuracy of 89.4%, sensitivity of 87.8%, specificity of 90.7%, and F1 score of 89.2%. Key features contributing to mortality prediction included Glasgow Coma Scale (GCS) scores, Braden Scale scores, blood urea nitrogen, age, red cell distribution width (RDW), bicarbonate, and lactate levels. Fairness analysis revealed that models trained on the top 30 most important features demonstrated lower DPD and EOD, suggesting reduced bias. Debiasing methods improved fairness in models trained with all features but had limited effects on models using the top 30 features.
Conclusions: ML models show strong potential for mortality prediction in critically ill hypertensive patients. Feature selection not only enhances interpretability and reduces computational complexity but may also contribute to improved model fairness. These findings support the integration of interpretable and equitable AI tools in critical care settings to assist with clinical decision-making.
Keywords:

Hypertension, Mortality Prediction, Machine Learning, Deep Learning, Fairness in AI,SHAP

37

Title: Telehealth Utilization and Patient Experiences: The Role of Social Determinants of Health Among Individuals with Hypertension and Diabetes

Author list: Haoxin Chen, Will Simmons, Malak Hashish and Jiancheng Ye

Abstract: Objective: To evaluate the utilization patterns, effectiveness, and patient satisfaction of telehealth services among individuals with hypertension and/or diabetes, and to investigate the influence of social determinants of health (SDOH) on telehealth access and utilization in this population.
Methods: We conducted a cross-sectional analysis using data from the 2022 Health Information National Trends Survey (HINTS 6) by the National Cancer Institute. The study sample included 3,009 respondents with self-reported diabetes, hypertension, or both conditions. Telehealth usage was assessed through 14 survey questions, and participant characteristics were analyzed using sociodemographic, baseline health, and SDOH data.
Results: Of the 6,252 HINTS 6 survey respondents, 3,009 met the inclusion criteria. Significant sociodemographic differences were observed across the diabetes and/or hypertension groups. No significant differences were found in telehealth usage among the groups, with 43.9% of respondents utilizing telehealth in the past year. Common reasons for telehealth use included provider recommendation, convenience, and infection avoidance. Social determinants of health, such as food insecurity and transportation issues, were more prevalent among individuals with both conditions, though no significant differences in telehealth experiences were noted across
groups.
Conclusion: Telehealth shows potential for managing chronic conditions like hypertension and diabetes, demonstrating substantial adoption and universal accessibility. However, disparities influenced by SDOH highlight the need for targeted interventions to ensure equitable access. Addressing privacy concerns, leveraging healthcare providers' recommendations, and tackling SDOH barriers are crucial for fostering wider telehealth adoption and improving outcomes. Future research should focus on the long-term impacts of telehealth and further investigate SDOH factors to develop tailored interventions that enhance engagement and equitable access across diverse patient populations.

Keywords: Telehealth; Technology utilization; Social determinants of health; Hypertension; Diabetes; Multiple chronic conditions

43

Title: MetaphorPrompt2 - A Structure and Function Focused Approach for Extracting Causal Events from Biological Text

Author list: Parth Patel, Yu-Chiao Chiu, Yufei Hunag, and Jianqiu Zhang

Abstract: Biomedical literature is crucial for building knowledge graphs that explain disease mechanisms and guide drug discovery. However, even advanced large language models (LLMs) using in-context learning often misinterpret complex domain-specific causal statements or omit intermediary steps, resulting in incomplete pathway representations. MetaphorPrompt2 is motivated by cognitive theories of causal event representation and analogical reasoning. It improves molecular regulation pathway (MRP) extraction by emphasizing the structural relations and functional roles of biological entities rather than relying on surface-level grammar. This enables more effective metaphor construction and structural alignment between expert and general domains. The system integrates five components that collectively reduce parsing complexity and mitigate error propagation. MetaphorPrompt2 outperformed previous

approaches, achieving a 24% improvement in edge prediction F1 score over a previous method without analogical reasoning. Notably, it eliminated missed entity errors and reduced m6A-related initiator extraction failures by 72.2%. These advances support the construction of more comprehensive biomedical knowledge graphs and enhance causal reasoning in LLMs, potentially facilitating automated hypothesis generation and accelerating drug discovery. Our findings highlight the value of a structure and function-focused approach for extracting complex causal knowledge from scientific text.
Keywords: Analogical Reasoning; LLM; Molecular Regulation Pathways; Knowledge Graphs; MetaphorPrompt; MetaphorPrompt2; Causal Events; Prompt Engineering

Technology Session
October 10th
3:40 PM – 5:40 PM
Room: 106

Chairs: Yu-Chiao Chiu, Juexin Wang

Title: Boosting Pipeline Efficiency in Bioinformatics Through Snakemake
Author list: Shunian Xiang[1], Hua ke[1], Jingling Hou[1], Nihir Patel[1], Yaoqi Li[1], Haixin Shu[1], Si Chen[1], Yaping Feng[1]
Detailed Affiliations:
[1]Department of Bioinformatics, Admera Health, New Jersey, NJ, USA
Abstract: Given the complexity and diversity of modern bioinformatics (BI) analyses, automation has become an essential priority—particularly for biotechnology companies that manage high volumes of multi-species projects with custom client requirements. Automating BI workflows through the integration of scripts, pipelines, and AI-assisted systems enables a more streamlined, assembly line-like approach, improving scalability, analysis speed, reproducibility, and reducing both manual effort and human error. Ultimately, this allows researchers to focus more on scientific discovery in biology and medicine.
We present a modular bioinformatics workflow framework built with Snakemake, a powerful and scalable workflow management system. Our system supports a broad spectrum of next-generation sequencing (NGS) data types, including bulk RNA-seq, small RNA-seq, microRNA-seq, single-cell RNA-seq, spatial transcriptomics, ChIP-seq, ATAC-seq and so on. Each pipeline is composed of independent, reusable modules—for example, the RNA-seq workflow includes quality control, adapter trimming, genome alignment, quantification, differential expression analysis, and pathway enrichment—which can be flexibly assembled and automatically executed through a simple command-line interface.
The system can automatically assemble workflows by selecting and combining appropriate modules based on user input, allowing flexible customization without sacrificing automation. This design significantly reduces manual workload, shortens turnaround time, and adapts easily to diverse project requirements—enabling efficient, reproducible, and scalable bioinformatics analysis.

Keywords: snakemake, automation, bioinformatics, efficiency, bulk RNA-seq, small RNA-seq, microRNA-seq, single-cell RNA-seq, spatial transcriptomics, chip-seq, atac-seq

Title: Spatial Transcriptomics at Scale with Stereo-seq: Big Data for Impactful Science
Author list: Yongfu Wang
Detailed Affiliations:
Complete Genomics
Abstract: Stereo-seq, originated from DNBSEQTM technology, is the highest resolution spatial transcriptomics platform available today. With 0.5 μm resolution and chip sizes up to 13 cm × 13 cm, Stereo-seq enables precise molecular mapping across whole organs or large tissue sections—ideal for developmental biology, oncology, neuroscience, and cross-species studies. This open, species agnostic platform supports both Fresh Frozen and FFPE samples, integrates transcriptomics with proteomics or histology, and captures total RNA—including non-coding RNAs and microbiome content—from FFPE samples. Hundreds to thousands of terabytes of data have been generated, and the scientific community continues to develop new tools to mine these datasets in pursuit of answers to the secret of life. This seminar will highlight the exciting advancements Stereo-seq has brought to the forefront of scientific discovery.

Title Access the full richness of biological complexity with single cell and spatial multiomics from 10x Genomics
Author list: Nicole Jaymalin
Detailed Affiliations:
10x Genomics, Pleasanton, California, USA.]
Abstract: Developing treatments for complex diseases requires building a complete understanding of both disease and treatment-response mechanisms. As we navigate a century where transformative advances in biology will reshape the way we deliver human health, translational and clinical researchers need approaches that provide actionable insights that can, ultimately, be leveraged to improve how diseases are diagnosed and treated.
Join us to learn how single cell, spatial, and in situ innovations from 10x Genomics can help you push the boundaries of your translational and clinical research. Discover novel therapeutic targets, explore how therapeutics modulate disease-associated cell populations and states, gain insights into mechanisms governing therapeutic toxicity, and understand resistance mechanisms governed by transcriptomic and epigenetic remodeling. Enabling deeper insight into cancer, immunology, neuroscience, and immuno-oncology, 10x Genomics gives researchers the ability to see biology in new ways.
Keywords: Single cell, Spatial Transcriptomics, In Situ

Title: Directed Evolution of Molecular Enzymes Empowers NGS Library Preparation
Author list: Robin Song
Detailed Affiliations:
Yeasen Biotechnology Co., Ltd.
Abstract: Next-generation sequencing (NGS) has transformed genomics, transcriptomics, and precision medicine by enabling high-throughput, large-scale nucleic acid analysis. At the heart of this revolution

lies molecular enzyme evolution, which continuously drives improvements in the sensitivity, specificity, and efficiency of sequencing workflows. Through advanced protein engineering and directed evolution, novel enzymes are developed to overcome technical challenges in library preparation, amplification, and data quality, paving the way for faster, more accurate, and cost-effective solutions.
This presentation will explore how cutting-edge enzyme innovation is reshaping the future of genomics by empowering genome and transcriptome sequencing , methylation analysis, and ultra-low input detection. We will highlight the role of enzyme-optimized reagent kits in enhancing experimental robustness and reliability, accelerating research breakthroughs, and expanding applications. By combining enzyme engineering with innovative sequencing approaches, we aim to help promote the future of genomics and enable new frontiers in life science and healthcare.
Keywords: Next-Generation Sequencing (NGS), Molecular Enzyme Engineering, Directed Evolution, Library Preparation, Genomic study, Transcriptome Analysis

Title: Uncover Cellular Heterogeneity with Advanced Single Cell Multi-Omics Approaches
Author list: Julie Laliberte[1], Jing Zhou[1]
Detailed Affiliations:
[1]Singleron Biotechnologies Inc., USA
Abstract: Singleron Biotechnologies is a pioneering molecular diagnostics company dedicated to advancing clinical diagnostics, drug development, and health management through cutting-edge single-cell analysis technologies. Singleron provides comprehensive solutions for single-cell sequencing, offering both instrument-based workflows for in-lab use and full-service options through our global network of service laboratories.
What sets Singleron apart is our proprietary single-cell partitioning system. Our SCOPE-chip platform utilizes a gravity-driven micro-well microfluidics system for gentle and efficient cell partitioning and RNA barcoding. This approach ensures robust performance—even with fragile, rare cell types or even nuclei.
In addition to high-throughput single-cell RNA sequencing, Singleron has developed a suite of innovative multi-omics technologies to extract deeper insights from each cell:

- **DynaSCOPE** detects nascent RNA, capturing transcriptional dynamics and providing valuable temporal information—particularly useful in applications like drug screening.
- **MobiuSCOPE** enables full-length transcript sequencing, making it ideal for detecting splice variants and SNPs across entire transcripts—overcoming the limitations of conventional 3′ or 5′ RNA-seq approaches.
- **FocuSCOPE** enriches specific transcripts alongside whole transcriptome profiling, increasing sensitivity for targeted gene expression analysis.
- **ProMoSCOPE** simultaneously profiles cell surface glycans and transcriptomes—offering critical insights into cell–cell interactions and immune responses.
- **Scircle** captures full-length T-cell and B-cell receptor sequences along with the whole transcriptome, allowing researchers to identify immune clonotypes of interest in cancer, autoimmune diseases, and vaccine development.

To support data interpretation, Singleron offers advanced analysis tools and resources.
Our SynEcoSys platform hosts over 46 million single cells from 731+ datasets across multiple species, all uniformly processed and expertly annotated to enable reliable cross-study comparisons.
Keywords: Single cell sequencing, Multiomics, Tissue dissociation, Advanced Bioinformatics tools

Future Scientist in AI Session

August 3rd
2:30 PM – 5:30 PM
Room: 301

Chair:

Flash Talk Session

August 5th
1:30 PM – 4:50 PM
Room 301

Chairs: Zhifu Sun

45
Title: A Multimodal Vision Transformer Using Fundus and OCT Images for Interpretable Classifications of Diabetic Retinopathy
Author list: Shivum Telang and Wei Chen
Abstract: Diabetic Retinopathy (DR) is a leading cause of vision loss worldwide, requiring early detection to preserve sight. Limited access to physicians often leaves DR undiagnosed. To address this, AI models leverage lesion segmentation for interpretability, but manually annotating lesions is impractical for clinical use. Physicians also require models that explain why a classification was made, not just where lesions are located. Current models often rely on a single imaging modality and achieve limited effectiveness. This study introduces RetGEN, a self-supervised learning-based framework that enhances DR classification through a multimodal vision-transformer architecture with a multimodal contrastive loss function. By integrating OCT and fundus scans, RetGEN improves classification accuracy while providing explainable insights for ophthalmologists. For interpretability, the model generates paired Grad-CAM heatmaps showcasing neuron weights across OCT images, visually highlighting regions contributing to DR severity classification. Trained on 3,000 fundus images, 1,000 OCT images, and 125 paired images, RetGEN outperforms state-of-the-art models, delivering more accurate, interpretable, and clinically meaningful assessments of DR severity. This methodology addresses key limitations in current DR diagnostics, offering a practical and comprehensive tool for improving patient outcomes.
Keywords: Vision Transformer; Contrastive Loss Function; Convolutional Neural Networks; Global Average Pooling; Weight Matrices

54

Title: In Silico Design of a Population-Specific mRNA Vaccine Targeting MUC1 for Colorectal Cancer: Focus on Iranian HLA Diversity

Author list: Sara Farahbakhsh, Zarrin Minuchehr, Raha Mahdavi Karimi, Hanita Kouzegar, Atrin Tofighi, and Amitis Masumian

Abstract: Advances in mRNA vaccines offer new cancer immunotherapy strategies. We designed an Iranian population-specific mRNA vaccine targeting MUC1, overexpressed in colorectal cancer. Bioinformatics identified common HLA alleles (HLA-A*24:02, HLA-B*35:01, HLA-C*04:01) and immunodominant epitope SVSDVPFPF with strong binding and high immunogenicity. The optimized vaccine demonstrated high stability and 98.3% population coverage, highlighting its potential as a targeted immunotherapy for MUC1-associated cancers and a framework for developing similar vaccines.

Keywords: mRNA vaccine; MUC1; Colorectal cancer; Bioinformatics

28

Title: Abnormal ERV expression and its clinical relevance in colon cancer

Author list: Aditya Vijay Bhagwate, Jason Ding, William Taylor, John Kisiel and Zhifu Sun

Abstract: Background: Human endogenous retroviruses (HERVs or ERVs) are genomic sequences that have integrated into the human genome from ancestral exogenous retroviruses and account for nearly 9% of human DNA. Many ERVs are expressed during embryogenesis but are epigenetically silenced afterward. However, growing evidence suggests that the reactivation of certain ERVs may be associated with human disease development, progression, and patient outcomes such as cancers and autoimmune diseases. Most studies selected a subset of ERVs and comprehensive profiling of well-annotated ERVs in colon cancer is lacking. This study aims to perform comprehensive profiling of ERVs and their associations with clinical phenotypes of colon cancer.

Methods: Cell line RNA-sequencing data, seven from colorectal cancer (CRC) and one from monocytes with both total RNA and polyA library preparations, and one from normal colon epithelium from PolyA protocol, were downloaded from RNA Atlas (GSE138734). RNA sequencing data for colon adenocarcinomas (COAD) and adjacent normal tissues were downloaded from GDC TCGA (https://portal.gdc.cancer.gov/). After alignment, ERV expression was quantified against comprehensively compiled ERVs (3,320). ERV expression profiles were compared between sequencing protocols, cancer and normal cells, and matched tumor and normal tissue pairs. ERV enrichment was performed with their overlapping or closest protein coding genes. Unsupervised clustering was used to identify ERV expression defined tumor subtypes and their associations with clinical and other molecular features. ERV association with disease specific survival (DSS) was performed using Cox regression model or Kaplan-Meier curves.

Results: ERV expressions between PolyA and total RNA protocols were comparable where both showed a higher number of expressed ERVs in cancer cells. The increased ERV expression was even more dramatic in primary COAD tumor samples. The “reactivated” or highly expressed ERVs in COAD were mainly located in intergenic region or intronic region of protein coding genes or lncRNAs. Host or nearby genes of these up-expressed ERVs were significantly enriched in viral protein interactions with cytokine and cytokine receptors while the down expressed genes were enriched in vitamin and ascorbate metabolism. ERV expression defined tumor classes were significantly associated with tumor mutation burden (TMB) and immuno-phenotypes such as antigen processing and presenting machinery (APM) and tumor immune infiltration score (TIS). Survival analysis identified 152 ERVs to be independently associated with DDS and 51 of them were also differentially expressed between tumors and normal samples.

Conclusions: ERV abnormal up expression is common in CRC. The ERV defined subtypes are associated with tumor immunity and some individual ERVs are independently associated with patient outcomes. These findings provide further evidence abnormal ERV expression has clinical and treatment implications.
Keywords: Human endogenous retrovirus; ERV; colon cancer; tumor immune response; patient survival

15
Title: From Bench to Insight: Rapid Pathogen Genomic Surveillance Workflow for SARS-CoV-2 and Emerging Pathogens
Author list: Chelsea Zimmer, Selena McVay, Keely Starke, Kimily Hughley, Sara N Koenig and Venkat Sundar Gadepalli
Abstract: Clinical surveillance of infectious diseases caused by viruses, such as SARS-CoV-2, is important for effective intervention and preventing potential epidemics or pandemics. The frequent mutations of the SARS-CoV-2 genome, caused by its RNA nature and lack of proofreading mechanisms, allow it to adapt to its host organisms. This adaptation can lead to new strains or variants of the virus. Historically, next-generation sequencing techniques required complex chemistry and specialized training of laboratory technicians and other specialized personnel. However, with improvements in automation and nanotechnology, these inherently specialized methodologies have been simplified and are easily adapted by the novice user in a clinical molecular lab. Parallel improvements in sequencing technologies and decreased costs associated with whole genome sequencing resulted in a worldwide effort of sequencing viral genomes from patients identified for SARS-CoV-2 infection. Large-scale analysis on these sequence data is now feasible with new bioinformatics pipelines and various reporting tools. Such pipelines allow a clinical laboratory to perform surveillance sequencing workflows without requiring advanced technical expertise, creating endless prospects. The array of sequence data generated across the globe offers diverse opportunities to study SARS-CoV-2 evolutionary dynamics and serves as a foundation for different research questions in the future. To enhance data accessibility for various research and global surveillance projects, public data repositories have been developed. These publicly accessible repositories host diverse data from different countries, thereby assisting in determining regional variants or identifying emerging variants. Even though bioinformatics tools are rapidly developed for identifying mutations and variant reporting, they require some computational expertise. We have developed a COVID-19 mutational analysis pipeline using Workflow Description Language (WDL), which is open-source and combines various steps in an analysis workflow with human-readable syntax. Thus, users with minimal informatics background can easily adapt the workflow while creating a local data repository within their institution. The pipeline processes input FASTA files and quality control files from Ion Torrent S5, performs clade and variant assignment, integrates patient metadata, and stores the results into a REDCap database. Further, in the pursuit of tracking sample records and relevant metrics during a sequencing run, our team has innovated a REDCap-based data capture system. This user-friendly REDCap form records essential details of each sequencing run and stores it in a REDCap database along with patient demographics info. To further enhance the utility of our REDCap-based data capture system, we have developed an intuitive interactive dashboard. This interface seamlessly connects with the REDCap data sources, providing real-time monitoring, interactive visualization, and the ability to create a consolidated variant report. Our overall approach streamlines processes in managing complex genomic data and offers easy adaptation to empower other molecular labs.
Keywords: WDL Workflow; Clade; Lineage; S5 Ion torrent suite

35

Title: LoRA-BERT: a Natural Language Processing Model for Robust and Accurate Prediction of long non-coding RNAs
Author list: Nicholas Jeon, Paul de Figueiredo, Lamin Saidykhan, Xiaoning Qian and Byung-Jun Yoon
Abstract: Long non-coding RNAs (lncRNAs) serve as crucial regulators in numerous biological processes. Although they share some sequence features similar to messenger RNAs (mRNAs), lncRNAs perform entirely different roles, providing important new avenues for biological research. The emergence of next generation sequencing technologies has greatly advanced the detection and identification of lncRNA transcripts and deep learning-based approaches have been introduced to classify lncRNAs. Although these methods have significantly improved the efficiency of identifying lncRNAs, they often lack robustness, and the prediction accuracy tends to vary significantly depending on the quality of the transcript. To tackle this issue, we introduce LoRA-BERT, a novel lncRNA prediction algorithm built on the bidirectional encoder representation from transformers (BERT). Lora-BERT is designed to effectively capture information at the nucleotide level that is important for lncRNA classification, leading to more robust and accurate prediction outcomes that are not significantly affected by the quality of the input transcript. Through performance evaluations on comprehensive benchmarks, we demonstrate that LoRA-BERT outperforms existing schemes in terms of accuracy, efficiency, and robustness. Especially, unlike other methods, LoRA-BERT retains good predictive capability for partial transcripts, a critical feature that makes it applicable for reliable lncRNA prediction even when the sequencing depth is relatively low.
Keywords: Natural language processing; Bidirectional Encoder Representations from Transformers (BERT); long non-coding RNA (lncRNA); lncRNA prediction

36
Title: ICM-MD: Integrating TM-Specific Features and MD-Derived Structures for Accurate Prediction of Inter-Chain Contacts in Alpha-Helical Transmembrane Homodimers
Author list: Bander Almalki and Li Liao
Abstract: Characterizing the interactions of alpha-helical transmembrane homodimers at the residue level is crucial for understanding their structure and function. However, most computational tools designed for globular proteins fail to translate to transmembrane (TM) proteins, largely due to the unique environment of the membrane and the limited availability of high-resolution structural data. To address this challenge, we present a machine learning framework geared for TM homodimers. Our method integrates sequence-based and structure-based features to enhance inter-chain residue contact prediction in TM homodimers.We address the challenge of limited training data by utilizing structures derived from molecular dynamics (MD) simulations as surrogate ground truth. Our model leverages a simple yet effective feed-forward neural network, designed to enhance model's interpretability and scalability. Comparativeevaluation against state-of-the-art models, including DeepHomo1, DeepHomo2, Glinter, and DeepTMP, demonstrates that our method achieves superior performance. On a test set of eight alpha-helical TM homodimers, our model outperforms DeepHomo1 and Deep-Homo2 by 155.5% and 261.0% respectively, surpasses Glinter by 92.0%, and achieves 11.6% higher precision compared to DeepTMP in the mean top L ranking metric.
Keywords: Transmembrane Proteins;Alpha-Helical homoidmers;Dimerization;Inter-Chain contact;Machine Learning

24
Title: OmicsSankey: Crossing Reduction of Sankey Diagram on Omics Data

Author list: Shiying Li, Bowen Tan, Si Ouyang, Zhao Ling, Miaozhe Huo, Tongfei Shen, Jingwan Wang and Xikang Feng
Abstract: In bioinformatics, Sankey diagrams have been widely used to elucidate complex biological insights by visualizing gene expression patterns, microbial community dynamics, and cellular interactions. However, computational scalability remains a challenge for large-scale biological networks. In this work, we present OmicsSankey, a novel formulation of the layout optimization problem for Sankey Diagrams that employs eigen decomposition as a closed-form solution, addressing graph disconnection through a teleportation mechanism that enhances connectivity and stabilizes eigenvector solutions. Experimental results on synthetic datasets with varying layers and nodes validate the efficacy of OmicsSankey compared to state-of-the-art layout-optimizers. Improving the Sankey layouts for Cell Layers, BioSankey, and Sequence Flow further validates OmicsSankey in enhancing the interpretability of biological insights.
Keywords: Sankey diagram; Layout optimization; Omics data visualization; Microbiome visualization

40
Title: Multi-omic analysis integrating global transcriptional and post-transcriptional profiles reveals predominant role of post-transcriptional control in three human cell lines
Author list: Alexander Krohannon, Mansi Srivasta, Neel Sangani and Sarath Janga
Abstract: Gene expression is regulated through a complex interplay between transcriptional and post-transcriptional mechanisms, yet their relative contributions and relationships remain incompletely understood. In this study, we propose normalized metrics to quantify regulatory density at each of these two levels by integrating ATAC-seq, RNA-seq, and Protein Occupancy Profiling sequencing (POP-seq) data across three human cell lines - HEK293, HepG2, and K562, to assess gene-specific regulatory variations. This analysis revealed 3 distinct regulatory classes: predominantly post-transcriptionally regulated, predominantly transcriptionally regulated, and neutrally regulated genes. Using this metric, significant associations between regulatory strategies and gene properties were uncovered; with transcriptionally regulated genes exhibiting greater length, post-transcriptionally regulated genes displaying higher isoform diversity and expression levels, and specific transcript types showing consistent enrichment patterns across regulatory categories. Remarkably, 55.8% of genes maintained identical regulatory classification across the three cell lines examined, with functional pathway analysis demonstrating high conservation of regulatory-functional relationships despite different cellular origins. The strong correlation between transcriptional and post-transcriptional regulation suggests coordinated interaction rather than independent operation. This study provides a novel framework for understanding gene regulatory strategies and demonstrates that the relationship between gene properties and regulatory mechanisms represents a fundamental organizational principle that transcends cell-type specificity, with implications for understanding dysregulation in disease states.
Keywords: Systems Biology; Multi-omics; Gene Regulation; ATAC-Seq; POP-Seq

22.
Title: TCR Convergence as a Proxy for Tumor-Specific Immunity in HSV1-Positive rGBM Patients Treated with CAN-3110
Author list: Ayse Selen Yilmaz[1,2], Alexander Ling[4], Hiroshi Nakashima[4], Xiaokui Mo[1,3], E. Antonio Chiocca[4]
Detailed Affiliations:
[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA [2] Bioinformatics Shared Resources, James Comprehensive Cancer

Center, The Ohio State University; Columbus, OH, USA [3] Center for Biostatistics, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA [4] Harvey Cushing Neuro-oncology Laboratories, Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA, USA
Abstract: Despite limited understanding on the mechanisms and predictors of outcome for oncolytic viral therapies, our previous work demonstrates that HSV1-positive rGBM patients exhibit prolonged survival following CAN-3110 viral therapy [1]. We hypothesize that this survival benefit is T cell mediated. T cells play an important role in adaptive immunity by utilizing the T cell receptors (TCRs) on their surfaces to recognize a wide range of antigens. Antigen specific TCRs are essential for eliminating tumor cells, but they are difficult to identify. A possible proxy for measuring tumor-specific TCRs is "TCR convergence", the phenomenon where TCRs have identical CDR3 amino acid sequences but different DNA sequences. TCR convergence has been shown to be a novel prognostic marker for immunotherapy [2]. We investigated relationships between TCR convergence, HSV1 serology, and survival in 21 IDH wild-type rGBM patients who received CAN-3110 treatment. DNA from pre- and post-treatment PBMCs was sequenced to extract TCRβ sequences. Patients were stratified into high and low TCR convergence groups based on the convergent TCR count in each pre-treatment sample, using the mean as a cutoff, and classified as positive or negative based on their HSV1 serology before treatment. We found that HSV1 seropositive patients are more likely to exhibit higher TCR convergence (Fisher's test, p-value=0.0071). Although a strong correlation between convergent TCR count and overall survival (OS) was not identified, high TCR convergence group shows higher OS compared to the low TCR convergence group (393.1 vs. 277.5 days, t-test, p=0.08). Given the observed association between TCR convergence and HSV1 serostatus, we next examined the structural organization and antigen specificity of the T cell repertoire. Using the TCRosetta platform [3], we identified TCR clusters at the CDR3 amino acid level, constructed TCR interaction networks, and predicted high-confidence peptide targets from VDJdb in pre- and post-CAN-3110 samples. Notably, top TCR clusters before and after treatment shared conserved CDR3 motifs, suggesting their sustained role in mediating anti-tumor immune responses. Our findings with this limited number of patients may guide future investigations into the role of TCR convergence in immunotherapy and its potential as a prognostic marker.

54.

Title: Vritra: a streamlined pipeline for species-resolved functional profiling of target genes in microbiome data

Author list: Boyan Zhou1, Menghan Liu2, Lama Nazzal3, Huilin Li1#

Detailed Affiliations:

[1]Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY, USA; [2]Department of Biological Sciences, Columbia University in the City of New York, New York, NY, USA; [3] Department of Medicine, New York University School of Medicine, New York, NY, USA.

Abstract: Microbiome functional profiling tools have advanced our understanding of microbial pathways and gene families. Increasingly, researchers are examining specific gene sets—for example, the frc/oxc genes involved in oxalate degradation or the bai clusters responsible for bile acid metabolism. Such targeted studies demand not only precise abundance quantification but also accurate species attribution. Two challenges remain: 1) ambiguous protein annotations hinder construction of comprehensive yet specific target gene sets, and 2) mainstream workflows align reads to UniRef90 clusters and then infer species in a manner that lacks the standardized boundaries (e.g., the ≥95 % ANI cutoff in GTDB) used to delineate species.

To overcome these limitations, we present Vritra (Versatile Reads-identification with Impartial Taxonomic Refinement and Assignment), a flexible pipeline for targeted gene detection and species-level profiling in shotgun microbiome sequencing. Vritra comprises two modules:
1. Gene-specific database construction. Users supply a single seed sequence plus a curated set of related sequences (e.g., from UniProt or InterPro). Vritra then employs a label-propagation algorithm to expand this into a refined UniRef90 reference tailored to the target gene family.
2. Species-resolved read analysis. Each UniRef100 cluster in the custom UniRef90 set is first assigned to a species using GTDB-style boundaries. Raw reads are then mapped to the representative sequence of each UniRef100 cluster to generate species-level gene abundance profiles.
Because Vritra builds only the gene-centric UniRef90 subset, its reference database is orders of magnitude smaller than the full UniProt UniRef90, greatly improving computational efficiency. We applied Vritra to three gene families across two publicly available microbiome cohorts, demonstrating accurate abundance estimates, robust species assignment, and broad applicability to both metagenomic and metatranscriptomic data.
Keywords: Microbiome; metagenomics; targeted gene profiling; species-level resolution; functional analysis

66.
Title: Supervised and unsupervised classification with feature selection for single-cell RNAseq based on an artificial immune system.
Author list: Dawid Krawczyk[1,4], Maciej Pietrzak[2], Michał Seweryn[3,4]
Detailed Affiliations:
[1]University of Lodz Doctoral School of Exact and Natural Sciences, University of Lodz, Poland;[2]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210, USA;[3]Centre for Digital Biology and Biomedical Sciences, Faculty of Biology and Environmental Protection, University of Lodz, Poland; [4]Regional Digital Medicine Center, Copernicus Memorial Hospital and University of Lodz, Poland
Abstract: In this study, we present a tool scAIS which enables both supervised and unsupervised classification of cells based on the expression profiles from single-cell RNA-seq experiments. Our approach is an extension of the method proposed by Dudek in doi: 10.1109/TEVC.2011.2173580. The main novelty of scAIS is related to the feature selection part which is performed simultaneously with the main task of learning – either supervised or unsupervised classification. In principle, scAIS is based on two main steps: (1) selection of epitopes (combinations of features/genes) which best separates the points of interest in high dimensional subspaces and (2) estimation of local neighborhood of data points (or clusters) which defines the local structure in lower-dimensional subspaces. The main advantage of scAIS is the ability to perform feature selection and clustering without dimension reduction performed on the initial step of preprocessing which is known to bias the final outcomes of the clustering as well as differential expression analysis – please refer to Rafael's Irizarry's blogpost https://simplystatistics.org/posts/2024-12-23-biologists-stop-including-umap-plots-in-yourpapers. We compare scAIS to the recent as well as classical methods of machine learning used for classification of single-sell RNAseq data. To this aim we use a set of benchmarking datasets available through the github repository as well as R package provided by prof Martin Hemberg's lab. In the side-by-side comparisons, based on both the real world data as well as simulation studies, to the novel scMINER and the classical SC3 algorithms our scAIS achieves comparable sensitivity of cluster detection and at the same time retains higher specificity.

Keywords
single-cell RNA sequencing, clustering, feature selection, artificial immune system, machine learning

76.
Title: VaxLLM: An end-to-end framework leveraging a fine-tuned Large Language Model for automated vaccine annotation and database integration
Author list: Xingxian Li[1,2], Matthew Asato[1], YuPing Zheng[3], Joy Hu[1], Feng-Yu (Leo) Yeh[2], Zhigang Wang[4], Jie Zheng[2], Yongqun He[2]
Detailed Affiliations
[1]College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI, USA; [2]Unit for Laboratory Animal Medicine, Center for Computational Medicine and Bioinformatics, Department of Learning Health Science, University of Michigan Medical School, Ann Arbor, MI, USA; [3]Chinese University of Hong Kong, Shenzhen, Guangdong, China; [4]Department of Biomedical Engineering, Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing, China.
Abstract: Vaccines play a vital role in enhancing immune defense and preventing hosts against a wide range of diseases. However, vaccine annotation remains a labor-intensive task due to the ever-increasing volume of scientific literature. This study introduces the Vaccine Large Language Model (VaxLLM) framework to explore the application of Large Language Model (LLM) in automating the annotation of scientific literature and database integration on vaccines.
To develop VaxLLM, we first fine-tuned the Llama 3 model using the training data from VIOLIN vaccine knowledgebase and PubMed articles. The VIOLIN knowledgebase has so far included 4,708 vaccines for 217 pathogens or non-infectious diseases (e.g., cancer), which provides comprehensive vaccine information. The paper processing started with the automatic fetching of articles by literature mining from PubMed. The fine-tuned model was first used to classify the articles to filter the relevant articles containing specific information about vaccine development. If the article was classified as “yes”, the fine-tuned Llama 3 model then annotated the article, specifically capturing key vaccine properties such as vaccine platform, antigen, formulation, target host species, experimental methods, protocols, immune response, efficiency, and experiment results. The PubTator tool was also used as an integrative component to extract the biomedical entities such as genes, diseases, chemicals, and species. To increase the accuracy of the model output, we also developed an annotation tool using the GPT API to extract more details from the full-text manuscript, such as detailed immune response and vaccine development stage, , and other related properties. For greater accuracy, a data harmonizer website was developed to help experts validate the results of these outputs in a clear manner. The final validated output could be directly exported into a database format and incorporated into the VIOLIN database.
Using keyword search, 143 PubMed articles about Brucella vaccines from the year 2024 to 2025 were used as testing data. The results of the VaxLLM were reviewed manually. The VaxLLM system achieved a classification precision of 0.92, recall of 1.0, AUROC of 0.88, and F1-score of 0.95. The annotation accuracy is 97.9%, outperforming the baseline Llama 3 model by a significant margin. Through rapid retrieval, the VaxLLM system can help the database increase its capacity at an extremely fast pace, such as adding thousands of vaccine information from literature in a year. Such a method may also be utilized for other domains of literature annotation.
Keywords: Vaccine, Large Language Model (LLM), Fine-tuning, Llama 3, PubTator, PubMed

80.

Title: Multi-Trait Polygenic Risk Score of Hypertension and Diabetes is Associated with Alzheimer's Disease Risk across Multi-Ethnic Cohorts

Author list: Anisha Das[1], Badri N. Vardarajan[1], and Annie J. Lee[1]

Detailed Affiliations:

[1]Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA

Abstract: Background: Vascular risk factors such as hypertension and diabetes are associated with increased risk for Alzheimer's Disease (AD). These conditions frequently co-occur and may share genetic components, yet their combined genetic influence on AD – particularly across diverse populations – remains underexplored. Methods: We assessed the shared polygenic contributions of hypertension and diabetes to AD risk using both single- and multi-trait polygenic risk score (PRS) approaches. Trait-specific PRSs were computed using PRS-CSx, which integrated GWAS summary statistics from European, African American, and East Asian ancestry cohorts, along with ancestry-matched linkage disequilibrium (LD) reference panels from the 1000 Genomes Project. We then implemented a multi-trait PRS (mtPRS) approach to account for the genetic correlation between vascular traits and their combined contribution to AD risk. A total of 27,831 individuals from six multi-ethnic cohorts were analyzed, and ancestry-specific and meta-analyses across ethnic groups were conducted to evaluate associations with AD dementia. Results: The mtPRS demonstrated a significant association with increased AD risk across all ancestry groups. Ancestry-specific odds ratios (ORs) ranged from 1.18-2.32, with consistent strong effect sizes and high statistical significance. For instance, in Non-Hispanic Whites, PRS-CSx produced ORs upto 2.319 (p = 5.74E-17), 1.847 (p = 1.05E-18), 1.720 (p = 7.99E-09), and 1.655 (p = 9.33E-15) in cohorts NACC, ROSMAP, ADNI, and WHICAP, respectively. Similar patterns were observed in African American, Caribbean Hispanic, and Mexican American cohorts. The meta-analysis across all ethnic groups further confirmed the additive contribution of vascular genetic risk to AD susceptibility (OR = 2.324, p = 1.39E-71). Conclusion: Our multi-trait PRS approach, which incorporates genetically correlated vascular traits and accounts for LD structure, improves the detection of genetic risk for Alzheimer's disease across diverse populations. These findings support the contribution of vascular factors to AD susceptibility and demonstrate the value of integrating related traits in polygenic risk modeling.

Keywords: Alzheimer’s Disease, Multi-trait Polygenic Risk Score, Multi-ethnic

---

Poster Session I
August 3rd
11:30 AM – 1:30 PM
Room: First floor Atrium

Poster Session II
August 4th
5:20 PM – 6:20 PM
Room: First floor Atrium

CONFERENCE LOCATION

Pomerene Hall

1760 Neil Ave, Columbus, OH 43210

The Translational Data Analytics Institute (TDAI) is a community of 1000+ faculty, researchers, staff and students from 70+ disciplines working at the forefront of interdisciplinary, big data-enabled science, scholarship and creative expression, with an emphasis on discoveries, solutions and insights for the greater good.

Parking Information

For guests wanting to park onsite, 12th Avenue Garage (340 West 12th Avenue, Columbus, OH 43210) offers visitor parking with a $14.5 daily maximum ($8.75 off-peak Max). The 12th Avenue Garage entrance is off of 12th Avenue, which is a couple of blocks away from Pomerene Hall.

View Map

Airport Information

John Glenn Columbus International Airport: 10 miles from the Conference Location.

Hotel Information

The Blackwell Hotel

Map and Directions

Special Conference Room Rate: $172 + tax and fees, Double or King.

Reservation details: please use this link to make a reservation.

SPECIAL ACKNOWLEDGEMENTS

We thank many people who helped with the peer-review of the manuscripts submitted to the conference. We are grateful for the numerous volunteering help and support from many people. We thank the Managing Director Cathie Smith, Brandon Elmore and Amanda Jovanovich from TDAI for their support and coordination of the events.

MANY THANKS TO OUR SPONSORS!

10x Genomics was founded on the vision that this century will bring advances in biomedicine and transform the way we understand and treat disease. We deliver powerful, reliable tools that fuel scientific discoveries and drive exponential progress to master biology to advance human health. Our end-to-end single cell and spatial solutions include instruments, consumables, and intuitive software, letting you unravel highly intricate biological systems, while bringing into focus the details that matter most.

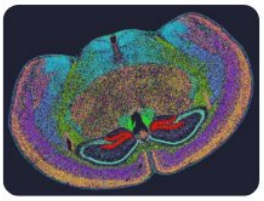
Complete GENOMICS™ STOmics

Complete Genomics is a pioneering life sciences company that provides novel, complete sequencing solutions, including sample/library preparation, lab automation, sequencing, and data analysis. The sequencing portfolio offers a comprehensive lineup of sequencers, ranging from low to high-throughput capacities, all powered by its proprietary DNBSEQ technology. Over 10,900 publications have been based on DNBSEQ technology across a wide range of applications.

Complete Genomics is the exclusive distributor for STOmics products in the US and Canada, featuring the revolutionary Stereo-seq technology. This powerful integration with DNBSEQ offers researchers comprehensive spatial transcriptomics capabilities, enabling detailed multi-omics investigations with unparalleled resolution and scale.

DNBSEQ Overview

Stereo-seq Overview



Learn more at completegenomics.com

For Research Use Only. Not for use in diagnostic procedures.

© Copyright 2025 Complete Genomics. All rights reserved. | 2904 Orchard Parkway, San Jose, CA 95134

Volume 118, October 2025 ISSN 1476-9271

ELSEVIER

Computational Biology and Chemistry

Editors: **Qin Ma, Donald Hamelberg**

www.elsevier.com/locate/cbac

Available online at www.sciencedirect.com

ScienceDirect

Olink’s mission is to accelerate proteomics together with the scientific community, to understand real-time biology and gain actionable insights into human health and disease. Our innovative solutions deliver highly sensitive and accurate protein quantification, giving scientists the power to investigate complex biological processes with precision.

One platform. Endless possibilities.

Explore up to 5,400 proteins with high specificity, transparent data, and the flexibility to answer any research question. Meet the next-generation proteomics platform trusted by the scientific community, from small academic research teams through to leading pharma companies.

Singleron

From single cell multi-omics to precision medicine

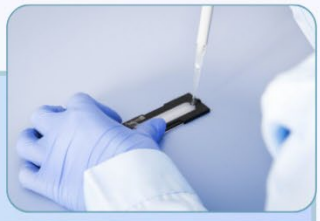
TISSUE PRESERVATION & DISSOCIATION

- Preserve tissue integrity for up to 72 hours - Maintain sample quality during transport or processing delays
- Automated and flexible - Adaptable programs to suit your needs
- Generate clean cell suspensions and isolate nuclei with ease

FLEXIBLE SINGLE CELL EXPERIMENTS

Manual workflow:
Instrument-free option for flexible, low- throughput needs

Automated workflow:
Fully automated cell partitioning & barcoding system generating sequencing-ready libraries



MULTIOMICS KITS

- (Full-length) transcriptome
- Full-length immune profiles
- Targeted variant detection
- Time-resolved transcriptomics
- Cell surface glycosylation
- Combined genome & transcriptome

GET IN TOUCH

Service lab: Singleron Michigan
333 Jackson plaza, Ann Arbor, MI 48103
+1 734-249-0883

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

OPEN ACCESS Impact Factor 4.4 CiteScore 9.3

Full peer review

Gold open access journal

Indexed in over 30 databases

Expert editorial board

We accept a range of article types

Research articles | Review articles | Mini reviews | Innovation reports
Short communications | Method articles | Database articles
So°ware/Web server articles | Perspectives | Editorials

CSBJ is composed of four sections and welcomes research in the following areas:

Functional and mechanistic understanding of how molecular components in a biological process work together, using computational methods

Editor-in-Chief: **Dr. Gianni Panagiotou**

New digital and automated technologies transforming health and care systems, with insights from real-world implementation in smart hospital settings

Editor-in-Chief: **Dr. Eleni Kaldoudi**

Advancing scientiÿc knowledge and technological innovation at the intersection of nanoscience, materials science, chemistry, physics, and biomedical engineering

Editor-in-Chief: **Dr. Antreas Afantitis**

Understanding biological systems that potentially harness quantum-mechanical processes and applying optics and photonic tools in quantum biology for biomedical and health sciences

Editor-in-Chief: **Dr. Youngchan Kim**

Find out more: https://www.csbj.org/

CSBJ is published by Elsevier on behalf of Research Networks

RESEARCH NETWORKS

ELSEVIER