ICIBM 2023

International Conference on Intelligent Biology and Medicine

July 16th-19th, 2023, Tampa, FL, USA

Hosted by:

The International Association for Intelligent Biology and Medicine (IAIBM)

2023 International Conference on
Intelligent Biology and Medicine
(ICIBM 2023)

July 16-19, 2023
Tampa, FL, USA

Hosted by:
The International Association for Intelligent Biology and Medicine (IAIBM)

Table of content

Welcome to ICIBM 2023!

On behalf of all our conference committees and organizers, we welcome you to the 2023 International Conference on Intelligent Biology and Medicine (ICIBM 2023). ICIBM is the official conference of The International Association for Intelligent Biology and Medicine (IAIBM, http://iaibm.org/), a non-profit organization whose mission is to promote intelligent biology and medical science, through member discussion, network communication, collaborations, and education.

The fields of bioinformatics, systems biology, and intelligent computing are continuing to evolve rapidly and have a strong impact on scientific research and medical innovations. With this in mind, we are proud to have built on the successes of previous years' conferences and provide a forum that fosters interdisciplinary research and discussions, educational opportunities, and collaborative efforts among these ever-growing and progressing fields.

This year, we have an exciting line-up for our keynote speakers, including Drs. Brooke Fridley, Bradley Malin, Jeffery Townsend, and Yidong Chen. Throughout the conference, we will also feature four eminent scholar speakers, Drs. Nancy Zhang, Peilin Jia, Lorin Crawford, and Tae Hyun Hwang. These researchers are world-renowned experts and we are privileged to host their talks at ICIBM 2023. We will also be hosting three workshops and two tutorials. In addition, talks will be given by faculty members, postdoctoral fellows, Ph.D. students and trainee-level awardees selected from outstanding manuscripts and abstracts. These researchers will showcase the innovative technologies and approaches that are the hallmark of our interdisciplinary fields.

Overall, we anticipate this year's program will be incredibly valuable to research, education, and innovation, and we hope you are as excited as we are to experience ICIBM 2023's program. We would like to extend our thanks to our sponsors for making this event possible, including the National Science Foundation, University of South Florida, University of Florida at Gainesville, Admera Health, 10x Genomics, Complete Genomics, One Florida^{+}: Clinical Research Network, Patterns: A Cell Press journal, and Computational and Structural Biotechnology Journal.

Last but not least, our sincerest thanks to members of all our ICIBM 2023 committees, and to our volunteers for their valuable efforts. Their dedication to making ICIBM 2023 a success is invaluable, and demonstrates the strength and commitment of our community.

On behalf of all of us, we hope that our hard work has provided a thought-provoking conference that fosters collaboration and innovation, and is enjoyable for all of our attendees. Thank you for attending ICIBM 2023. We look forward to your participation in all our conference has to offer!

Sincerely,

Xiaoming Liu, PhD
Program Co-Chair
Associate Professor,
College of Public Health Genomics Program, University of South Florida

Fuhai Li, PhD
Program Co-Chair,
Assistant Professor,
Washington University in St. Louis

Li Liu, MD
Program Co-Chair
Associate Professor,
Arizona State University

Kai Wang, PhD
General Co-Chair
Professor
Children's Hospital of Philadelphia

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

Jiang Bian, PhD
General Co-Chair
Professor,
University of Florida Health Center

ACKNOWLEDGEMENTS

General Chairs
Kai Wang, Children's Hospital of Philadelphia, USA
Zhongming Zhao, The University of Texas Health Science Center at Houston, USA
Jiang Bian, University of Florida, USA

Steering Committee
Kun Huang, Indiana University, USA
Yves Lussier, University of Utah, USA
Jason Moore, Cedars-Sinai Medical Center, USA
Jake Chen, University of Alabama at Birmingham, USA
Sudhir Kumar, Temple University, USA
Lawrence Hall, University of South Florida, USA

Program Chairs
Xiaoming Liu, University of South Florida, USA
Fuhai Li, Washington University, USA
Li Liu, Arizona State University, USA

Program Committee
James Cai, Texas A&M University
Xiao Chang, Children's Hospital of Philadelphia
Kaifu Chen, Harvard University
Bin Chen, Michigan State University
Junjie Chen, Harbin Institute of Technology, Shenzhen
Jianlin Cheng, University of Missouri Columbia
Feng Cheng, University of South Florida
Kyuhong Cho, Indiana State University
Zechen Chong, University of Alabama at Birmingham
Yulin Dai, Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston
Babu Guda, University of Nebraska Medical Center
Yan Guo, Vanderbilt University
Jun-Tao Guo, UNC Charlotte
Matthew Hayes, Xavier University of Louisiana
Eric Ho, Lafayette College
Gangqing Hu, West Virginia University
Ruifeng Hu, UTHealth-Houston
Tao Huang, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
Frank Huang, Cincinnati Children's Hospital Medical Center
Zhi-Liang Ji, Xiamen University
Peilin Jia, UTHealth
Limin Jiang, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Jeff Kinne, Indiana State University
Fuhai Li, Washington University in St. Louis
Shuai Cheng Li, City University of Hong Kong
Aimin Li, Xi'an University of Technology
Li Liao, University of Delaware
Li Liu, Arizona State University
Xiaoming Liu, University of South Florida
Qian Liu, University of Nevada, Las Vegas
Bingqiang Liu, shandong University
Shuang Luan, University of New Mexico
Qin Ma, The Ohio State University
Tianle Ma, Oakland University
Vijay Mago, Lakehead University
Mirjana Maletic-Savatic, Baylor College of Medicine
Sayaka Miura, Temple University
Hong Qin, University of Tennessee at Chattanooga
Yufeng Shen, Columbia University
Li Shen, University of Pennsylvania
Yang Shen, Texas A&M University
Xinghua Shi, Temple University
Yijun Sun, SUNY Buffalo
Fengzhu Sun, University of Southern California
Zhifu Sun, Mayo Clinic
Jijun Tang, University of South Carolina
Haixu Tang, Indiana University Bloomington
Ladda Thiamwong, University of Central Florida
Manabu Torii, Kaiser Permanente
Jun Wan, Indiana University School of Medicine
Kai Wang, University of Pennsylvania
Jiayin Wang, Xi'an Jiaotong University
Junbai Wang, Radium Hospital
Chaochun Wei, Shanghai Jiao Tong University
Junfeng Xia, Institute of Physical Science and Information Technology, Anhui University
Jingfa Xiao, Beijing Institute of Genomics
Lei Xie, City University of New York
Jinchuan Xing, Rutgers University
Min Xu, Carnegie Mellon University
Yu Xue, Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
Jingwen Yan, Indiana University-Purdue University Indianapolis
Wei Zhang, University of Central Florida
Fan Zhang, University of North Texas Health Science Center
Zhongming Zhao, University of Texas Health Science Center at Houston
Jim Zheng, School of Biomedical Informatics, University Texas Health Science Houston
Jiang F. Zhong, Loma Linda University
Yunyun Zhou, Children's Hospital of Philadelphia

Publication Committee
Kaifu Chen, Harvard University, USA
Gangqing Hu, West Virginia University, USA

Workshop/Tutorial Committee
Jun Wan, Indiana University
Yi Guo, University of Florida

Publicity Committee
Rays Jiang, University of South Florida, USA
Guimin Gao, U Chicago, USA
Mattia Prosperi, University of Florida, USA

Award Committee
Li Chen, University of Florida, USA
Mingyao Li, University of Pennsylvania, USA

Trainee Committee
Keith L Sanders, University of Texas, USA
Sumarga Sah Tyagi, University of South Florida

Local Organization Committee
Feng Cheng, University of South Florida, USA
Yicheng Tu, University of South Florida, USA

SCHEDULE

Sunday, July 16th

<table>
<tr><td>11:00 AM-
6:00 PM</td><td colspan="4">Registration</td></tr>
<tr><td colspan="5">CONCURRENT WORKSHOPS</td></tr>
<tr><td colspan="2">Room: St. Petersburg I</td><td colspan="2">Room: St. Petersburg II, III</td><td>Room: Williams/Demens</td></tr>
<tr><td colspan="2">Technology Session

Chair: Zhongming Zhao</td><td colspan="2">Workshop on Applications of AI in Translational Research

Chairs: Zhe He, Rui Yin</td><td>Flash Talks

Chairs: Kaixiong Ye, Chengqi Wang</td></tr>
<tr><td>2:00 PM-
2:30 PM</td><td>Unveiling the complexity of breast cancer through advanced analysis of FFPE tissue: single cell, spatial, and in situ mapping of the tumor microenvironment

Dr. Ryan Mote (10x Genomics)</td><td>2:00 PM-
2:20 PM</td><td>Adaptive graph model deciphers spatial cellular communications

Dr. Qianqian Song (Wake Forest University)</td><td>Shared genetic basis informs the roles of polyunsaturated fatty acids in brain disorders

Huifang Xu, Yitang Sun, Michael Francis, Claire Cheng, Nitya Modulla, Kaixiong Ye

Common Genetic Variants are Associated with Plasma and Skin Carotenoid Metabolism in Ethinically Diverse US Populations

Yixing Han, Savannah Mwesigwa, Melissa N. Laska, Stephanie B. Jilcott Pitts, Nancy E. Moran, Neil A. Hanchard</td></tr>
<tr><td>2:30 PM-
3:00 PM</td><td>StereoCell: A bioinformatics tool enables accurate single-cell segmentation for spatial transcriptomics dataset

Dr. Shan Yang (Complete Genomics)</td><td>2:20 PM-
2:40 PM</td><td>Leveraging the power of genomics to facilitate the diagnosis of undiagnosed diseases with machine learning models

Dr. Rui Yin (University of Florida)</td><td>Cross-analysis between P. falciparum Var expression with host immunothrombosis markers to better define pediatric cerebral malaria phenotypes.

Iset Vera, Thomas Keller, Anne Kessler, Visopo Harawa, Wilson L. Mandala, Stephen J. Rogerson, Terrie E. Taylor, Karl B. Seydel, and Kami Kim</td></tr>
</table>

				Unveiling Gene Interactions in Alzheimer's Disease by Integrating Genetic and Epigenetic Data with a Network-Based Approach Keith Sanders, Astrid M Manuel, Andi Liu, Boyan Leng, Xiangning Chen, Zhongming Zhao
3:00 PM-3:30 PM	**A scRNA-seq cell type identifying method based on human curated cell marker database and empirical knowledge** Dr. Yaping Feng (Admera Health)	2:40 PM-3:00 PM	**Better Acute Kidney Injury Prediction and Risk Factor Analysis with Personalized Transfer Learning** Dr. Mei Liu (University of Florida)	**MalariaSED: a deep learning framework to decipher the regulatory contributions of noncoding variants in malaria parasites** Chengqi Wang, Yibo Dong, Jenna Oberstaller, Chang Li, Min Zhang, Justin Gibbons, Camilla Valente Pires, Lei Zhu, Rays H.Y. Jiang, Kami Kim, Jun Miao, Thomas D. Otto, Liwang Cui, John H. Adams, Xiaoming Liu
				Enhancing DNA Sequence Matching and Ranking through Deep Learning-Based Alignment-Free Model Sumarga K. Sah Tyagi, Minh Pham, Yicheng Tu
3:30 PM-3:45 PM	***Break***	3:00 PM-3:20 PM	**Harnessing Explainable, Equitable, and Actionable AI to Improve Health** Dr. Zhe He (Florida State University)	**3D genome reveals intratumor heterogeneity in Glioblastoma** Qixuan Wang, Juan Wang, Qiushi Jin, Mark W. Youngblood, Lena Ann Stasiak, Ye Hou, Yu Luan, Radhika Mathur, Joseph F. Costello, Feng Yue
				Integrated Spatial Multi-omics Analysis Based on MALDI Data

				Xin Ma, Cameron Shedlock, Harrison Clarke, Roberto Ribas, Terrymar Medina, Tara R. Hawkinson, Shannon Keohane, Craig W. Vander Kooi, Matthew S. Gentry, Li Chen, Ramon Sun
		3:20 PM-3:45 PM	***Break***	
3:45 PM-4:15 PM	**Generating real-world evidence using OneFlorida+ clinical research consortium** Dr. Yi Guo (University of Florida)	3:45 PM-4:05 PM	**Constructing a Large-Scale Biomedical Knowledge Graph and Its Applications in Drug Discovery** (via Zoom) Dr. Jinfeng Zhang (Florida State University)	**Multimodal machine learning combining image and textual data to predict rare genetic disorders** (recorded video) Da Wu, Jingye Yang, Kai Wang
				A multimodal neuroimaging-based risk score for Alzheimer's disease by combining clinical and large N>37000 population data Elaheh Zendehrouh, Mohammad SE. Sendi, Vince D. Calhoun
4:15 PM - 4:45 PM	**Unraveling the Challenges of Genomic Sequencing and Computational Analysis: Introducing the Genomics Sequencing Core and Computational Core** Drs Min Zhang & Bi Zhao (University of South Florida)	4:05 PM-4:25 PM	**Translational Pharmaco-informatics research** Dr. Lai Wei (Ohio State University)	**Developing an Accurate and Interpretable Risk-Based Model for Lung Cancer Screening** Piyawan Conahan, Lary Robinson, Haley Tolbert, Margaret M Byrne, Lee Green, Yi Luo
				In silico Improvement of Highly Protective Antimalarial Antibodies Mateo Reveiz, Andrew Schaub, Young Do Kwon, Prabhanshu Tripathi, Azza Idris, Amarendra Pegu,Lais Da Silva Pereira, Patience

				Kiyuka, Myungjin Lee, Tracy Liu, Chen-Hsiang Shen, Baoshan Zhang, Yongping Yang, Peter D. Kwong, Reda Rawi
		4:25 PM-4:45 PM	**Using Explainable Machine Learning Models to Predict CAR T-Cell Therapy Response with Longitudinal Patient Report Outcomes** Dr. Yi Luo (Moffitt Cancer Center)	**Comprehensive Investigation of Active Learning Strategies for Anti-Cancer Drug Response Prediction** Priyanka Vasanthakumari, Yitan Zhu, Thomas Brettin, Alexander Partin, Maulik Shukla, and Rick L. Stevens
				Bioinformatics and machine learning based identification of potential oxidative stress and glucose metabolism diagnostic Biomarkers in Alzheimer disease Sidra Aslam, Fatima Noor, Thomas G. Beach, Geidy E. Serrano

Monday, July 17th

<table>
<tr><td>8:00 AM- 6:30 PM</td><td colspan="3">Registration</td></tr>
<tr><td colspan="4">CONCURRENT WORKSHOPS/TUTORIALS</td></tr>
<tr><td>Room: St. Petersburg I</td><td colspan="2">Room: St. Petersburg II, III</td><td>Room: Williams/Demens</td></tr>
<tr><td rowspan="8">Tutorial on Collection and Analyses of Long-Read Transcriptome and Epitranscriptome Data

Organizer: Kin Fai Au (University of Michigan)

9:30 AM-12:00 PM

Requirement: bring your laptop</td><td colspan="2">Workshop on Microbiome Data Analysis

Chair: Qunfeng Dong</td><td rowspan="8">Artificial Intelligence on Big Data: Promise for Early-stage Trainees

Chairs: Yufang Jin, Chi Zhang, Yongsheng Bai

Trainee presentations (10 min)

9:00 AM – 11:10 AM

Feasibility of a 3D Convolutional Neural Network for the Diagnosis of Alzheimer’s Disease using Brain PET Scans (Troy Zhang)

Comparisons of Coronavirus Spike Proteins and the Mutation Effects on Virus-Host Interaction (Crystal Teng)

Identification of Key Biomarkers Associated with Ductal Breast Cancer in Spatial Transcriptomics Data (Ellie Xi)

Assessing the Clinical Significance Identification Capability of DNA Language Models: A Study of Enformer's Performance on Disease-Causing Variants in Human Cis-Regulatory Elements (Rain Hou)

Characterization of oncogenes and tumor suppressor genes with onco-microRNAs and tumor suppressor microRNAs (Claire Shen)</td></tr>
<tr><td>9:30 AM-9:50 AM</td><td>Artificial intelligence-based identification of microbes in cancer

Dr. Noam Auslander (The Wistar Institute)</td></tr>
<tr><td>9:50 AM-10:10 AM</td><td>Metagenomic read classification using deep learning models (via Zoom)

Dr. Ying Zhang (University of Rhode Island)</td></tr>
<tr><td>10:10 AM-10:30 AM</td><td>De Novo Identification of Contaminants in Low Microbial Biomass Microbiomes

Dr. Yunxi Liu (Rice University)</td></tr>
<tr><td>10:30 AM-10:50 AM</td><td>Multiscale adaptive differential abundance analysis in microbial compositional data

Dr. Shulei Wang (University of Illinois Urbana-Champaign)</td></tr>
<tr><td>10:50 AM-11:00 AM</td><td>Break</td></tr>
<tr><td>11:00 AM-11:20 AM</td><td>A novel microbial causal mediation analytic pipeline for</td></tr>
</table>

		investigating microbiome's mediating role in disease and health disparity (via Zoom) Dr. Huilin Li (New York University)	**FLUXestimator: a webserver for predicting metabolic flux and variations using transcriptomics data** (Alex Lu) **Identifying relationships between cellular topology and gene expression in spatial transcriptomics of breast cancer tissues** (Isabella Wu) **Pan-cancer analysis of metabolic shifts via flux estimation analysis** (Kevin Hu) **Temporal Phenotyping for Transitional Disease Progress: an application to cardiovascular diseases and neurological diseases** (Andy Wang) **The artificial intelligence analysis of single-cell transcriptomes highlights the high heterogeneity in bladder cancer** (Xilin Wei) **Tissue Domains Identification using Spatial Transcriptomics Data** (Emily Wei) **Adaptive Deep Inference with Collaborative Architecture for IoT** (Alejandro Villanueva) **Analysis of thermal images for Nearby Animal Behavior using Deep learning architectures for enhancing vehicle safety** (Eleni Avlonitis)
	11:20 AM-11:40 AM	**Incorporating metabolic activity, taxonomy and community structure to improve microbiome-based predictive models for host phenotype prediction** Dr. Mahsa Monshizadeh (Indiana University)	
	11:40 AM-12:00 PM	**Exploring the Male Urethral Microbiome: A Community Ecology Approach Based on the Neutral Theory**	**Panel Discussion** 11:10 AM – 12:00 PM

		Dr. Xiang Gao (Loyola University Chicago)	
12:00 PM - 1:30 PM	***Lunch Break***		
1:30 PM – 1:40 PM	**Opening Remarks (St. Petersburg II, III)**		
CONCURRENT SESSIONS/WORKSHOP			
	Room: St. Petersburg I	**Room: St. Petersburg II, III**	**Room: Williams/Demens**
	Workshop on Prompt Bioinformatics – Application of ChatGPT and Large Language Models Chair: Gangqing Hu	**Genomics, Transcriptomics, Proteomics and Epigenomics I** Chairs: Qin Ma, Xiaojing Wang	**Medical Informatics, Public Health Informatics and Pharmacoinformatics I** Chairs: Mei Liu, Satish Mahadevan Srinivasan
1:40 PM - 2:00 PM	**Prompt Bioinformatics with Chatbots** Dr. Gangqing Hu (West Virginia University)	**Eminent Scholar** **Nancy Zhang (University of Pennsylvania)** **Title: Signal recovery in single cell data integration**	**The association between nonalcoholic fatty liver disease (NAFLD) status and physical exam or biochemical parameters** Weiru Han, Tianrui Zhu, Zhengli Tang, Robert Morris, Kun Bu, Fang Wang, Lin Fan, Weijian Wang, Yiming Hao, Yiqin Wang and Feng Cheng
2:00 PM - 2:20 PM	**PROMPT BIOINFO. CASE STUDY: Intra-tumor Evolutionary Inference** Dr. Sayaka Miura (Temple University)	**Mitigating Heterogeneity Effects in Microbiome-based Quantitative Phenotype Prediction: A Comprehensive Workflow for Integrating Multiple Studies with Batch Normalization** Yilin Gao and Fengzhu Sun	**Behavioral and demographic profiles of HIV contact networks in Florida** Yiyang Liu, Christina Parisi, Rebecca Frisk-Hoffman, Marco Salemi, Diego Viteri, Mattia Prosperi and Simone Marini
2:20 PM - 2:40 PM	**Leveraging Stand-alone RNA-Seq Data for Novel lncRNA Identification and Annotation: A Prompt**	**Comprehensive Cross Cancer Analyses Reveal Mutational Signature Cancer Specificity** Rui Xin, Limin Jiang, Hui Yu, Jijun Tang and Yan Guo	**Association between ABCG1/TCF7L2 and type 2 diabetes mellitus：An intervention trial based case-control study**

	Bioinformatics Case Study Dr. Chan Zhou (University of Massachusetts)		Yinxia Su, Xiangtao Liu, Conghui Hui, and Hua Yao
2:40 PM - 3:00 PM	**Exploring ChatGPT's Ability to Generate Novel Algorithms in Bioinformatics** Dr. Li Liu (Arizona State University)	**A Weighted Two-stage Sequence Alignment Framework to Identify DNA Motifs from ChIP-exo Data** Yang Li, Yizhong Wang, Cankun Wang, Anne Fennell, Anjun Ma, Jing Jiang, Zhaoqian Liu, Qin Ma and Bingqiang Liu	**Smoothing spline analysis of variance models: A new tool for the analysis of accelerometer data** Rui Xie, Lulu Chen, Joon-Hyuk Park, Jeffrey Stout and Ladda Thiamwong
3:00 PM - 3:15 PM	***Coffee/Tea Break***		
3:15 PM - 3:35 PM	**Enhanced Gene Interaction Analysis and Pathway Reconstruction through Iterative Prompt Refinement by ChatGPT** Yibo Chen, Mihail Popescu, Dong Xu	**A mouse-specific model to detect genes under selection in tumors** Hai Chen, Jingmin Shu and Li Liu	**Exploring Drug-drug Interaction Information from PubMed using Association Rules** Kun Bu, Weiru Han, Robert Morris and Feng Cheng
3:35 PM - 3:55 PM	**Ensemble BERT for Medication Event Classification on Electronic Health Records (EHRs)** Shouvon Sarker, Xishuang Dong and Lijun Qian	**A machine learning pipeline to detect open chromatin regions from cfDNA sequencing data** Yuxin Liu, Yuqian Liu, Xiaoyan Zhu, Jiayi Ren, Xin Lai, Xuanping Zhang and Jiayin Wang	**Pan-cancer mutational signature surveys correlated cancer racial disparities with geospatial environmental exposures, and viral infections** Judy Bai, Katherine Ma, Shangyang Xia, Richard Geng, Limin Jiang, Hui Yu, Xi Gong, Shuguang Leng and Yan Guo

3:55 PM - 4:15 PM	**Enhancing Phenotype Recognition in Clinical Notes Using Large Language Models: PhenoBCBERT and PhenoGPT** Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou and Kai Wang	**Detection of viral infection in cell lines using ViralCellDetector** Rama Shankar, Shreya Paithankar, Suchir Gupta and Bin Chen	**Characterizing Diseases using Genetic and Clinical Variables: A Data Analytics Approach** Madhuri Gollapalli, Harsh Anand and Satish Mahadevan Srinivasan
4:15 PM - 4:35 PM	**Flash Talk: PROMPT BIOINFO. CASE STUDY: Shotgun Metagenomic Data Analysis** Zhu Xing, Qiyun Zhu **Flash Talk: Cancer Comprehend Annotation – a pipeline for cancer phenotype and clinical extraction** Thanh Duong, Phillip Szepietowski, Thanh Thieu	**A comprehensive benchmark of transcriptomic biomarkers for immune checkpoint blockades** (recorded video) Hongen Kang, Xiuli Zhu, Ying Cui, Zhuang Xiong, Wenting Zong, Yiming Bao and Peilin Jia	**The Association between Warfarin usage and International normalized ratio increase: Systematic analysis of FDA Adverse Event Reporting System (FAERS)** Robert Morris, Matthew Bruckner, Milagros Salcedo, Nicole Zapata Aponte, Alfredo Suarez Garcia, Megan Todd, Weiru Han, Kun Bu, Feng Cheng and Rachel Webb
4:35 PM - 4:50 PM	***Coffee/Tea Break***		
4:50 PM - 5:30 PM	**Keynote Lecture (Room: St. Petersburg II, III)** **Yidong Chen, Ph.D. (University of Texas Health Science Center at San Antonio)** **Title: Learning cellular responses to genetic and chemical perturbations of cancer**		
5:30 PM - 5:40 PM	***Break***		
5:40 PM - 6:40 PM	**Poster Session (Room: St. Petersburg II, III)** Poster size: 3’ (width) x 4’ (height), portrait form		
7:00 PM – 9:00 PM	**Reception (Ford Garage at 200 1st Avenue S)**		

Tuesday, July 18th

8:00 AM-5:30 PM	**Registration**		
8:30 AM - 9:10 AM	**Keynote Lecture (St. Petersburg II, III)** **Bradley Malin, Ph.D. (Vanderbilt University)** **Title: Building Ethically Viable Biomedical Data Science Environments**		
9:10 AM - 9:20 AM	***Break for parallel sessions***		
CONCURRENT SESSIONS			
	Room: St. Petersburg I	**Room: St. Petersburg II, III**	**Room: Williams/Demens**
	Special Session on Dynamics of Transcriptional Regulation Towards Single Cell, Single Molecular, and Spatial Omics Session Chair: Kaifu Chen	**Computational Methods for Aging and Brain Research** Chairs: Shaolei Teng, Guogen Shan	**Genomics, Transcriptomics, Proteomics and Epigenomics II** Chairs: Renzhi Cao, Jing Wang
9:30 AM - 9:50 AM	**Uncovering Disease-Associated Novel lncRNAs: A Computational Perspective** Dr. Chan Zhou (University of Massachusetts)	**Eminent Scholar** **Peilin Jia (Beijing Institute of Genomics, China)** **Title: Deep learning approaches for accurate drug response imputation (via Zoom)**	**AlphaCluster: Coevolutionary driven residue-residue interaction models enable quantifiable clustering analysis of de novo variants to enhance predictions of pathogenicity** Joseph Obiajulu, Ranger Kuang, Lesi He, Guojie Zhong, Jacob Hagen, Chang Shu, Wendy Chung and Yufeng Shen
9:50 AM - 10:10 AM	**Data-driven and AI-empowered systems biology** Dr. Chi Zhang (Indiana University)	**Clustering Alzheimer's Disease Subtypes via Similarity Learning and Graph Diffusion**	**Mutation Density Analyses on Long Noncoding RNA Reveal Comparable Patterns to Protein-Coding RNA and Prognostic Value**

		Tianyi Wei, Shu Yang, Davoud Ataee Tarzanagh, Jingxuan Bao, Jia Xu, Patryk Orzechowski, Joost B. Wagenaar, Qi Long and Li Shen	Chaoyi Troy Zhang, Hui Yu, Yongsheng Bai and Yan Guo
10:10 AM - 10:30 AM	**Deep learning reveals cellular state transition** Dr. Guangyu Wang (Houston Methodist Research Institute)	**Machine Learning Analysis for Studying Aging-Associated Hearing Loss** Safa Shubbar	**Systematic assessment of small RNA profiling in human extracellular vesicles** Jing Wang, Hua-chang Chen, Quanhu Sheng, Renee Dawson, Robert J. Coffey, James G. Patton, Alissa M. Weaver, Yu Shyr, Qi Liu
10:30 AM - 10:50 AM	**MEBOCOST: Metabolite-mediated Cell Communication Modeling by Single Cell Transcriptome** Dr. Kaifu Chen (Harvard Medical School)	**Vagus nerve stimulation and blood pressure modulate neuronal activity in the periventricular cerebellum** Maria Alejandra Gonzalez-Gonzalez	**Flash Talk: Genomic disparities between cancers in adolescent and young adults and in older adults** Xiaojing Wang, Anne-Marie Langevin, Peter Houghton, Siyuan Zheng **Flash Talk: TSSr: an R package for comprehensive analyses of TSS sequencing data** Zhaolian Lu, Keenan Berry, Zhenbin Hu, Yu Zhan, Tae-Hyuk Ahn, Zhenguo Lin
10:50 AM - 11:05 AM	***Coffee/Tea Break***		

11:05 AM - 11:25 AM	**Enhancing Cell-Type Identification in Single-Cell RNA-seq Data with Interpretable Deep Learning** Dr. Liang Chen (University of Southern California)	**A Machine Learning Based Multiple Imputation Method for the Health and Aging Brain Study-Health Disparities** (via Zoom) Fan Zhang, Melissa Petersen, Leigh Johnson, James Hall, Raymond Palmer and Sid O'Bryant	**The genetic regulation of the biogenesis of human isomiRs** Guanglong Jiang, Jill L. Reiter, Chuanpeng Dong, Yue Wang, Fang Fang, Zhaoyang Jiang and Yunlong Liu
11:25 AM - 11:45 AM	**The Whole is More Than the Parts: Decoding Synergistic Networks of Multiple Non-coding Variants Linked to Cancer Risk** Dr. Xueqiu Lin (Stanford University)	**Structure-learning-based causal comorbidities mining from UK biobank: an exploratory study for Alzheimer's disease** Yiheng Pan, Pingjian Ding, Zhenxiang Gao and Rong Xu	**SynthQA - Hierarchical Machine Learning-based Protein Quality Assessment** Mikhail Korovnik, Sheng Wang, Junyong Zhu, Kyle Hippe, Jie Hou, Dong Si, Kiyomi Kishaba and Renzhi Cao
11:45 AM - 12:05 PM	**Flash Talk: Building the Human Ensemble Cell Atlas and Learning the Underlying Unified Coordinate System** Xuegong Zhang	**Flash Talk: Decentralization of Brain age Estimation with Structural Magnetic Resonance Imaging Data** Sunitha Basodi, Rajikha Raja, Bhaskar Ray, Harshvardhan Gazula, Jingyu Liu, Eric Verner and Vince D. Calhoun	**Characterizing protein structural features of alternative splicing and isoforms using AlphaFold 2** Yuntao Yang, Yuhan Xie, Zhao Li, Chiamaka Diala, Meer Ali, Rongbin Li, Yi Xu, Sayed-Rzgar Hosseini, Erfei Bi, Hongyu Zhao and Wenjin Zheng
	Flash Talk: Deep Transfer Learning of Cancer Drug Responses by Integrating Bulk and Single-cell RNA-seq data Junyi Chen, Xiaoying Wang, Anjun Ma, Qi-En Wang, Bingqiang Liu, Lang Li, Dong Xu, Qin Ma	**Flash Talk: An integrative study to identify the link between dysregulated intercellular signaling and genetic variants in Alzheimer's disease** Andi Liu, Xiaoyang Li, Brisa S Fernandes, Yulin Dai, Zhongming Zhao	
12:05 PM - 1:40 PM	***Lunch Break***		
1:40 PM - 2:20 PM	**Keynote Lecture (St. Petersburg II, III)**		

	Brooke Fridley, Ph.D. (Moffitt Cancer Center) **Title: Decoding Kidney Cancer: Analytical Strategies for Unveiling the Tumor Immune Microenvironment using Spatial Transcriptomics**		
2:20 PM - 2:30 PM	***Break for parallel sessions***		
CONCURRENT SESSIONS			
	Room: St. Petersburg I	**Room: St. Petersburg II, III**	**Room: Williams/Demens**
	Single Cell Omics Data Modeling and Analysis Chairs: Qianqian Song, Guangyu Wang	**Machine Learning/Deep Learning in Biomedical Research I** Chairs: Jinchuan Xing, Qian Liu	**Medical Informatics, Public Health Informatics and Pharmacoinformatics II** Chairs: Yi Guo, Lijun Cheng
2:30 PM - 2:50 PM	**Improving cellular phylogenies through integrated use of mutation order and optimality principles** Sayaka Miura, Tenzin Dolker, Maxwell Sanderford and Sudhir Kumar	**Eminent Scholar** **Lorin Crawford (Microsoft Research and Brown University)** **Title: Interpretable Probabilistic Models to Identify Multi-scale Enrichment in Complex Traits**	**Predicting COVID-19 Severity of Emergency Room Patients using Chest X-ray Images** Jonathan Stubblefield and Xiuzhen Huang
2:50 PM - 3:10 PM	**Gradient boosting reveals spatially diverse cholesterol gene signatures in colon cancer** Xiuxiu Yang, Justin Couetil, Debolina Chatterjee, Valerie Ardon, Jie Zhang, Kun Huang and Travis Johnson	**Revealing the impact of genomic alterations on cancer cell signaling with an interpretable deep learning model** Shuangxia Ren, Jonathan Young, Xinghua Lu and Lujia Chen	**Quantifying the Growth of Glioblastoma Tumors Using Multimodal MRI Brain Images** Anisha Das, Shengxian Ding, Rongjie Liu and Chao Huang
3:10 PM - 3:30 PM	**scDemultiplex: An iterative beta-binomial model-based method for accurate demultiplexing with hashtag oligos** Li-Ching Huang, Lindsey Stolze, Alexander Gelbard,	**DeepCORE: An interpretable multi-view deep neural network model to detect co-operative regulatory elements** Pramod Bharadwaj Chandrashekar, Hai Chen,	**Comparing the risk of deep vein thrombosis of two combined oral contraceptives: norethindrone/ethinyl estradiol and drospirenone/ethinyl estradiol**

	Yu Shyr, Qi Liu and Quanhu Sheng	Matthew Lee, Navid Ahmadinejad and Li Liu	Jennifer Stalas, Robert Morris, Kun Bu, Kevin von Bargen, Rebekah Largmann, Kathryn Sanford, Jacob Vandeventer, Weiru Han and Feng Cheng
3:30 PM - 3:50 PM	**Decoding ecosystem heterogeneity and transcriptional regulation characteristics of multi-subtype renal cell carcinoma** (recorded video) Kailong Xu, Jie Liu, Heng Yang, Lixin Ma, Gang Dou and Yang Wang	**A novel interpretable k-hop graph attention network model of integrative omics data analysis to infer target-specific core signaling pathways** Ruoying Yuan, Jiarui Feng, Heming Zhang, Yixin Chen, Philip Payne and Fuhai Li	**An In-silico Study of Antisense Oligonucleotide Antibiotics** (via Zoom) Erica Chen and Eric Ho
3:50 PM - 4:00 PM	***Coffee/Tea Break***		
4:00 PM - 4:20 PM	**Improving cell type identification at single-cell level** Mostafa Malmir, Jinyan Li, Anita Omo-Okhuasuy, Umar Jamil, Yidong Chen and Yufang Jin	**Proformer-based Ensemble Learning for Gene Expression Prediction** Lucy Nwosu, Xiangfang Li, Seungchan Kim, Lijun Qian and Xishuang Dong	**Reducing the Data for Radiation Cancer Therapy Quality Assurance** Maryam Albuainin, Richard Shaw and Shuang Luan
4:20 PM – 4:40 PM	**Osteogenic Differentiation Potential of Mesenchymal Stem Cells using Single Cell Multiomic Analysis** Duojiao Chen, Xiaona Chu, Hongyu Gao, Patrick McGuire, Xuhong Yu, Xiaoling Xuei, Yichen Liu, Sheng Liu, Jill Reiter, Jun Wan, Yunlong Liu and Yue Wang	**DeepDecon accurately estimates cancer cell fractions in bulk RNA-seq data** Jiawei Huang, Yuxuan Du, Andres Stucky, Jiang F. Zhong and Fengzhu Sun	**Exploring How Healthcare Organizations Use Twitter: A Discourse Analysis** Aditya Singhal and Vijay Mago
4:40 PM – 5:00 PM	**Do Single-cell Hi-C Data Follow A Power Law Distribution?**	**Accurate prediction of functional effect of single missense variants with deep learning**	**The Association Between Bradycardia and the Use of Remdesivir**

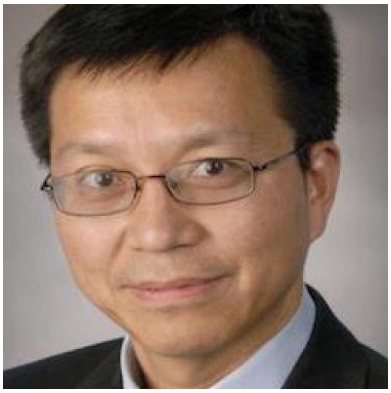
	Bin Zhao, Patrick Shen and Lu Liu	Houssemeddine Derbel, Zhongming Zhao and Qian Liu	Gibret Umeukeje, Robert Morris, Weiru Han, Kun Bu, Jin Wei, Ruisheng Liu and Feng Cheng
5:00 PM - 5:30 PM	***Award presentation*** (Williams/Demens)		
6:00 PM – 9:00 PM	**Banquet** (St. Petersburg II, III)		

Wednesday, July 19th

8:00 AM-12:00 PM	**Registration**		
8:30 AM - 9:10 AM	**Keynote Lecture (St. Petersburg II, III)** **Jeffrey Townsend, Ph.D. (Yale University)** **Title: Dismantling the Coarse Paradigm of Cancer 'Drivers' and 'Passengers'**		
9:10 AM - 9:20 AM	***Break for parallel sessions***		
CONCURRENT SESSIONS/TUTORIAL			
	Room: St. Petersburg I	**Room: St. Petersburg II, III**	**Room: Williams/Demens**
	Special Topics on Genomics and Translational Bioinformatics Chairs: Ece Uzun, Shulan Tian **AMIA GenTBI Working Group** 9:30 AM – 11:10 PM	**Cancer Informatics and Network Biology** Chairs: Noam Auslander, Xinna Zhang	**Machine Learning/Deep Learning in Biomedical Research II** Chairs: Xiao Fan, Yijie Wang
9:30 AM - 9:50 AM	**Chromatin-mediated transcriptional dysregulation in T-cell Prolymphocytic Leukemia** Dr. Huihuang Yan (Mayo Clinic)	**Eminent Scholar** **Tae Hyun Hwang (Mayo Clinic)** **Title: Harnessing single cell and spatial genomics with machine learning and AI to develop biomarker and therapeutic strategies for immuno and cellular therapy in cancer**	**Metastatic cancer expression generator (MetGen): A generative contrastive learning framework for metastatic cancer generation** Zhentao Liu, Yu-Chiao Chiu, Yidong Chen and Yufei Huang
9:50 AM - 10:10 AM	**Decoding Genomic Variations with Variant Graph Craft: A User-Friendly Tool for VCF Analysis** Dr. Alper Uzun (Brown University)	**CoMatch: a transfer learning model connecting in vivo finding to outcome prediction to distinguish prognostic/predictive biomarkers in breast cancer**	**Neural relational inference optimization to analyze enzyme allosteric interactions in regular enzymes** (recorded video)

		Abhishek Majumdar, Aida Yazdanparast, Huanmei Wu, Lang Li and Lijun Cheng	Shuang Wang, Yan Wang, Yi He, Xuhong Zhang, Weiwei Han and Juexin Wang
10:10 AM - 10:30 AM	**Genomics and Artificial Intelligence in Clinical Care** Dr. Nephi Walton (Intermountain Healthcare)	**Identifying Significantly Perturbed Subnetworks in Cancer Using Multiple Protein-Protein Interaction Networks** Le Yang, Runpu Chen, Thomas Melendy, Steve Goodison and Yijun Sun	**CCLHunter: an efficient toolkit for cancer cell line authentication** (recorded video) Congfan Bu, Xinchang Zheng, Jialin Mai, Zhi Nie, Jingyao Zeng, Qiheng Qian, Tianyi Xu, Yanling Sun, Yiming Bao and Jingfa Xiao
10:30 AM - 10:50 AM	**Unified somatic calling and machine learning-based classification enhance the discovery of clonal hematopoiesis of indeterminate potential** Dr. Shulan Tian (Mayo Clinic)	**Integrating and interpreting multi-omics data via novel k-hop graph neural network models to uncover core disease signaling pathways in medulloblastoma** Zitian Tang, Jiarui Feng, Yixin Chen, Philip Payne and Fuhai Li	**Seizure prediction based on deep learning driven by nonlinear dynamics** Wei Xiaoyan, Zhen Zhang and Yi Zhou
10:50 AM - 11:00 AM	**Unveiling the Hidden Web of Protein-Protein Interactions in Cancer and Subtypes** Dr. Ece Uzun (Brown University) 10:50 AM – 11:10 AM	***Coffee/Tea Break***	
11:00 AM - 11:20 AM		**scGEM: unveiling the nested tree-structured gene co-expressing modules in single-cell transcriptome data** Han Zhang, Xinghua Lu, Binfeng Lu and Lujia Chen	**A Transformer-Based Deep Learning Approach for Fairly Predicting Post-Liver Transplant Risk Factors** Can Li, Xiaoqian Jiang and Kai Zhang
11:20 AM - 11:40 AM	**Unlocking the Power of Single-Cell Gene Expression Analysis with SCGEATOOL: A Hands-On Tutorial for Non-Programmers and Machine Learning Experts**	**Repurposing drugs for Group3 and Group4 medulloblastoma subtypes by inhibiting novel common core signaling targets** Fuhai Li, William Buchser, Clifford Luke, Di Huang, Maxene Ilgan and Joshua Rubin	**DRLCOMPLEX: Reconstruction of Protein Quaternary Structures Using Deep Reinforcement Learning** Elham Soltanikazemi, Raj Roy, Farhan Quadir, Nabin Giri, Alex

	Organizer: James Cai (Texas A&M University) 11:20 AM – 12:30 AM (Note: free Matlab license available for software installation)		Morehead and Jianlin Cheng
11:40 AM - 12:00 PM		**Prediction of prognosis, immunotherapy and chemotherapy with an immune-related risk score model in endometrial cancer** (recorded video) Wei Wei, Zhenting Huang, Bo Ye, Xiaoling Mu, Jing Qiao, Peng Zhao, Yuehang Jiang, Jingxian Wu and Xiaohui Zhan	**Pan-cancer drug response prediction through tumor decomposition by cancer cell lines** Yu-Ching Hsu, Yu-Chiao Chiu, Tzu-Pin Lu, Tzu-Hung Hsiao and Yidong Chen
12:00 PM – 12:20 PM		**Flash Talk: A massive proteogenomic screen identifies thousands of novel peptides from the human "dark" proteome** Xiaolong Cao, Siqi Sun, Jinchuan Xing	**Flash Talk: HiC4D: Forecasting spatiotemporal Hi-C data with residual ConvLSTM** Tong Liu, Zheng Wang
		Flash Talk: Mutated processes predict immune checkpoint inhibitor therapy benefit in metastatic melanoma Andrew Patterson, Noam Auslander	**Flash Talk: Integrating Hydrogen Bonding Information into Graph Neural Networks for Protein Structure Classification** Yi-Shan Lan, Tsung-Yi Ho
12:20 PM - 12:30 PM	**Closing Remarks** (St. Petersburg II, III)		



Keynote Speaker
Yidong Chen, Ph.D.
Monday, July 17, 2023
4:50 PM – 5:30 PM
St. Petersburg II, III

Dr. Yidong Chen received his B.S./M.S. degrees in Electrical Engineering from Fudan University, Shanghai, China, and Ph.D. in Imaging Science from Rochester Institute of Technology, Rochester, NY. He has been with Hewlett Packard Co as a Research Engineer before he joined National Institutes of Health (NIH) in 1996. At NIH, Dr. Chen joined microarray technology development effort at National Human Genome Research Institute (NHGRI), as a Special Expert, Staff Scientist, and later Associate Investigator for microarray image, statistical analysis, and bioinformatics. From 2006-2008 he joined the Genetics branch at National Cancer Institute (NCI) as a staff scientist. During the 13-year period with NHGRI and NCI, he has contributed about 90 peer-reviewed publications and book chapters. Currently, Chen Lab works on computational biology and bioinformatics and focuses on developing computational solutions and statistical modeling to bridge between quantitative science and the basic biology and translational research within Greehey Children's Cancer Research Institute and around UT Health San Antonio.

Title: Learning cellular responses to genetic and chemical perturbations of cancer

Advances in high-throughput technologies have revolutionized the research of pharmacogenomics in cancer, which studies drug response based on a patient's genetic makeup. Empowered by the ability to discover knowledge from big data, deep learning (DL) has emerged to be one of the best techniques to characterize and learn from rapidly accumulating pharmacogenomics data. In this presentation, we will discuss the DL models to capture genetic and pharmacologic perturbations that induce similar effects on cell viability or molecular changes. We have proposed several deep learning models that integrate heterogeneous datasets of genome-wide CRISPR loss-of-function screens, high-throughput pharmacologic screens, and base-line genomic data that facilitate the characterization, discovery, and prediction of connections between omics, response to genetic and chemical perturbation of cancer samples, and drive towards the realization of precision oncology. Several new developments in handling proteomic data, as well as single-cell RNA-seq data to expand the prediction of perturbation response, will be discussed. These new developmental efforts will further push the deep learning methods to the biologists who can quickly search, visualize and connect their own dataset and knowledge to rapidly evolving genetic and chemical perturbation data sets. Taken together, our novel DL models can be used to characterize tumor types and subtypes, assist in the optimal treatment selection for cancer patients, and explore the mechanism of actions of oncological compounds in the realization of precision oncology.

Keynote Speaker
Bradley A. Malin, Ph.D.
Tuesday, July 18, 2023
8:30 AM – 9:10 AM
St. Petersburg II, III

Brad Malin, Ph.D., is Vice Chair for Research Affairs in the Department of Biomedical Informatics at Vanderbilt University Medical Center and the Accenture Professor of Biomedical Informatics, Biostatistics, and Computer Science at Vanderbilt University. In 2016, he founded GetPreCiSe, one of the National Human Genome Research Institute Centers of Excellence on Ethical, Legal, and Social Implications Research, which investigates multidisciplinary approaches to maintaining privacy in the collection, use, and sharing of genetic information. In 2021, he became one of the PIs of the NIH's Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) initiative and in 2022, he became the lead PI for the Ethics and Trustworthy AI core of the NIH's Bridge2AI Center. In addition, has served as co-chair of the Committee on Access, Privacy, and Security (CAPS) for the All of Us Research Program since its inception and is an appointed member of the Board of Scientific Counselors of the National Center for Health Statistics of the Centers for Disease Control and Prevention (CDC). Among various honors, he is an elected fellow of the National Academy of Medicine and was a recipient of the Presidential Early Career Award for Scientists and Engineers (PECASE) from the White House.

Title: Building Ethically Viable Biomedical Data Science Environments

You simply cannot perform data science investigations without access to data! And yet, the collection, use, and subsequent dissemination of biomedical data raises many ethical questions and societal quandaries that have the potential to thwart such activities. The goal of this presentation is to discuss how ethical reasoning and computational decision making can be embedded into data-driven activities in biomedicine and ultimately maximize social good. This presentation will touch on issues of trust, algorithmic fairness, data privacy, and public policy. Along the way, we will review real case studies to understand how things have gone wrong in the past, how they have gone right, and what the future may hold, particularly in the face of rapidly advancing artificial intelligence techniques! This talk will draw upon examples from large-scale data driven environments created at the local level, such as the de-identified electronic medical record (EMR) and bioreposity developed at Vanderbilt University Medical Center, the Electronic Medical Record and Genomics (eMERGE) consortium of the NIH, and the All of Us Research Program, which is working towards the creation of a database of EMRs, genome sequences, and mHealth records from one million Americans.

Keynote Speaker
Brooke Fridley, Ph.D.
Tuesday, July 18, 2023
1:40 PM – 2:20 PM
St. Petersburg II, III

Brooke L. Fridley is Professor, Senior Member and Chair of the Department of Biostatistics and Bioinformatics at Moffitt Cancer Center in Tampa, FL. She is also the Scientific Director for Moffitt Cancer Center's Biostatistics and Bioinformatics Shared Resource. Prior to joining Moffitt Cancer Center, Dr. Fridley was at the University of Kansas Medical Center and the Mayo Clinic in Rochester, MN. At the University of Kansas Medical Center, she was Director of the Biostatistics and Informatics Shared Resource for the NCI designated University of Kansas Cancer Center and Site Director for the Kansas-INBRE Bioinformatics Core. Dr. Fridley received her BS in mathematics from Truman State University and her MS and PhD in statistics from Iowa State University. Her research focus is in the areas of statistical genomics, molecular epidemiology of cancer, cancer genomics and pharmacogenomics. She has extensive experience as a collaborating statistician, particularly in the design and analysis of studies involving multiple types of 'omic data. Recently, she has been particularly involved in studies of ovarian cancer, development of integrative analysis methods for study with multi-omic data, and development of methods and software for the analysis of the spatial architecture of the tumor microenvironment. She has over 280 publications, has been awarded 5 NIH grants, is the Scientific Director for Moffitt's Lung Metabolism P01 Data Science Core, and is MPI of a NIH/NCI T32 training grant in cancer biology and data science.

Title: Decoding Kidney Cancer: Analytical Strategies for Unveiling the Tumor Immune Microenvironment using Spatial Transcriptomics

Immunotherapy (IO) treatments for clear cell renal cell carcinoma (ccRCC) have been developed to leverage patients' immune system against malignant tissues and have yielded improvements in patient survival. However, many tumors become resistant following IO. Hence, understanding how IO changes the tumor immune microenvironment (TIME) leading to resistance in ccRCC is critical. Exploring the heterogeneity in the TIME is key to understanding the prognosis of cancer and development of effective cancer therapies. Single-cell RNA sequencing (scRNAseq) is a powerful tool for studying the TIME at the single-cell level, however, the spatial context in which cells occur is not preserved in this technology. In contrast, the use of spatially resolved transcriptomics holds promise in the understanding of the spatial contexture of the TIME because of its power to capture gene expression profiles and the locations of individual cells within the larger tissue architecture. Recently, commercialization of spatially resolved transcriptomics (ST) technologies have allowed researchers to get a better understanding of spatial architecture of the TME. In this presentation, an overview is presented for some of the analytical and visualization approaches available for ST data. In particular, the adaptation and application of methods from the fields of spatial and ecological statistics are discussed. To demonstrate their application, some of the findings from a series of ccRCC tissue samples are presented. The data set represents a cohort of patients with primary treatment naive ccRCC and resistant ccRCC tissue after IO treatment. Tumor tissue samples were assayed using the NanoString's single-cell CosMx ST platform. An overview of the methods to assess the significance of the tissue architecture for IO outcomes is presented, beginning with the identification of ccRCC cell populations and culminating with the inference of potential cell interactions from the perspective of spatial co-localization and co-expression analyses.

Keynote Speaker
Jeffery Townsend, Ph.D.
Wednesday, July 19, 2023
8:30 AM – 9:10 AM
St. Petersburg II, III

Jeffrey Townsend is a biostatistician and evolutionary biologist. He earned a Bachelor of Science in biology from Brown University in 1994, taught in primary and secondary education for three years, then earned a PhD in organismic and evolutionary biology from Harvard University in 2002. He is currently the Elihu Professor of Biostatistics and Professor of Ecology and Evolutionary Biology at the Yale School of Public Health at Yale University, where he serves as Co-Director of the Genetics, Genomics, and Epigenetics Research Program at the Yale Cancer Center. He is the Co-Chair of the American Association for Cancer Research Cancer Evolution Working Group.

Title: Dismantling the Coarse Paradigm of Cancer 'Drivers' and 'Passengers'

The terms “driver” and “passenger” in cancer served a valuable purpose early in cancer genomics, when the primary goal of genomic inference was to identify genes to add to a scarce list of genes that could be investigated molecularly for potential therapeutic targeting and drug discovery. However, as larger numbers of genes have been identified with distinct mutational, proliferative, or cell survival roles in oncogenesis, and prioritization has become an essential component of investigation, I would like to challenge the utility of the prevailing notions of 'drivers' and 'passengers'. Employing advances in computational biology, I’ll demonstrate how a comprehensive quantification of the context-specific strength of selection on a variant that accounts for underlying mutation rates surpasses the performance of commonly used mutation effect metrics like SIFT and PolyPhen in predicting experimentally-based COSMIC Tiers of variant relevance to oncogenesis. This more accurate understanding of the impact and functional consequences of mutations within cancer driver genes can be extended to analyses yielding cancer stage-specific estimates of selection on cancer drivers. By capturing the dynamic nature of cancer progression, these estimates offer valuable insights into the selective forces acting at different stages of the disease, vastly enhancing our understanding of tumor evolution. This understanding can be further expanded to incorporate pairwise somatic genetic interactions, supplanting common concepts of mutual exclusivity and co-occurrence of driver genes with quantification of antagonistic and synergistic selective epistasis on variants. Revealing the intricate interplay between the cancer effects of somatic mutations sheds light on the complexity of gene functional relationships and provides a more nuanced understanding of cancer development. In turn, estimation of three-way, four-way, and five-way selective epistasis enables estimation of the changing selective effects on somatic genetic evolutionary trajectories in cancer. By visualizing the diverse paths that cancer genomes can take, we gain a deeper understanding of the mutational landscapes and their implications for tumor progression and treatment strategies. Lastly, we delve into the quantification of selection for genetic changes conferring cancer resistance to therapy. By elucidating the selective pressures driving therapeutic resistance, we reveal crucial insights into the development of effective treatment approaches. This challenge to the coarse paradigm of describing cancer-related genes with the binary terms 'drivers' and 'passengers' offers improved site-specific mutation effect metrics, better stage-specific understanding of gene action and interaction, and the potential to develop precision therapeutic strategies that anticipate, forestall, or eliminate the evolution of resistance.

Eminent Scholar Talk
Nancy Zhang, Ph.D.
Monday, July 17, 2023
1:40 PM – 2:00 PM
St. Petersburg II, III

Dr. Zhang is a Ge Li and Ning Zhao Professor of Statistics in The Wharton School at the University of Pennsylvania. Her research focuses primarily on the development of statistical methods and computational algorithms for the analysis of genomic data. She has made contributions to copy number and structural variant detection and to intra-tumor genetic heterogeneity modeling, and recently she has made myriad methodological contributions to the analysis of single-cell sequencing data. In Statistics, she has made contributions to change-point analysis, variable selection, and model selection. Dr. Zhang obtained her Ph.D. in Statistics in 2005 from Stanford University. After one year of postdoctoral training at the University of California, Berkeley, she returned to the Department of Statistics at Stanford University as Assistant Professor in 2006. She received the Sloan Fellowship in 2011, and formally moved to University of Pennsylvania in 2012. She was awarded the Medallion Lectureship by the Institute of Mathematical Statistics in 2021. Her work has been funded by grants from the NSF and NIH. At Penn, she is a member of the Graduate Group in Genomics and Computational Biology, and currently serves as the Vice Dean of the Wharton Doctoral Program.

Title: Signal recovery in single cell data integration

Data integration to align cells across batches has become a cornerstone of population-level single cell studies, critically affecting downstream analyses. Yet, how much signal is erased from data during integration? Currently, there are no guidelines for when biological signals are separable from batch effects, and thus, studies usually take a black-box, trial-and-error attitude towards data integration. I will show evidence that current paradigms for single cell data integration are unnecessarily aggressive, removing biologically meaningful variation. To remedy this, I will present a novel statistical model and computationally scalable algorithm, CellANOVA, to recover biological signal that is lost during single cell data integration. CellANOVA utilizes a "pool-of-controls" design concept, common in population-level single cell studies, to separate unwanted variation from biological variation of interest. When applied with existing integration methods, CellANOVA allows the preservation of subtle biological signals and substantially corrects the data distortion introduced by integration. Further, CellANOVA explicitly estimates cell- and gene-specific batch effect terms which can be used to identify the cell types and pathways exhibiting the largest batch variations, providing clarity as to which biological signals can be recovered.

Eminent Scholar Talk
Peilin Jia, Ph.D.
Tuesday, July 18, 2023
9:30 AM – 9:50 AM
St. Petersburg II, III

Peilin Jia, PhD, obtained her Ph.D. degree in bioinformatics in Shanghai Institutes for Biological Sciences, 2008. She received her Postdoctoral training in Virginia CommonWealth University and Vanderbilt University. She was an Assistant Professor (research-track) in Vanderbilt University (2012-2016) and an Assistant Professor (tenure-track) in the University of Texas Health Science Center at Houston (2016-2020). She is currently a Professor in Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation since 2021. Dr. Jia has a broad interest in bioinformatics, genomics, precision medicine, and machine learning and has co-authored 170 papers in these areas (cited by >9,000 times, H-index = 49). Recently, Dr. Jia developed several computational methods using the vast amount of cancer genomics data to identify driver mutations and mutational processes, curate tumor suppressor genes, and predict drug response based on baseline gene expression profile, with the overarching goal of translating computational discoveries to improve human health.

Title: Deep learning approaches for accurate drug response imputation

Prediction of therapy response has been a major challenge in cancer precision medicine due to the extensive tumor heterogeneity. While many studies have been conducted to identify signature genes or biomarkers to infer drug sensitivity or resistance, few have investigated the significant roles of transcriptome contexts in shaping the eventual treatment outcome. In this study, we developed a deep variational autoencoder (VAE) model followed by Elastic Net (VAEN) to predict the response for a total of 251 compounds using pre-treatment transcriptome data. To enable interpretability, we further developed a VAE-based model with a mask layer to integrate a priori knowledge of biological pathways and named it mVAEN. We validated these methods using multiple independent datasets assessing multiple compounds for their predictive capability of drug response, survival outcome, and cell line status. In addition, we implemented several embedding-based methods for the task of drug response prediction and benchmarked their performance. We demonstrated that using the embedding vectors could accurately impute drug response, outperform standard signature-gene-based or linear-compression-based approaches, and appropriately control the overfitting problem. Finally, we developed DrVAEN (https://bioinfo.uth.edu/drvaen), a user-friendly and easy-accessible web-server for drug-response prediction and model comparison for broad use in cancer research, method evaluation, and drug development.

Eminent Scholar Talk
Lorin Crawford, Ph.D.
Tuesday, July 18, 2023
2:30 PM – 2:50 PM
St. Petersburg II, III

Lorin Crawford is a Principal Researcher at Microsoft Research New England, and he holds a faculty position as an Associate Professor of Biostatistics at Brown University. His lab develops machine learning algorithms and statistical tools to understand how non-additive variation plays a role in complex traits and contributes to disease in diverse human populations. Some of his most recent work has landed me a place on Forbes 30 Under 30 list and recognition as a member of The Root 100 Most Influential African Americans in 2019. He has also been awarded an Alfred P. Sloan Research Fellowship, a David & Lucile Packard Foundation Fellowship for Science and Engineering, and a COPSS Emerging Leader Award. Prior to joining both MSR and Brown, Dr. Crawford received his PhD from the Department of Statistical Science at Duke University and a Bachelor of Science degree in Mathematics from Clark Atlanta University.

Title: Interpretable Probabilistic Models to Identify Multi-scale Enrichment in Complex Traits

A common goal in genome-wide association (GWA) studies is to characterize the relationship between genotypic and phenotypic variation. Linear models are widely used tools in GWA analyses, in part, because they provide significance measures which detail how individual single nucleotide polymorphisms (SNPs) are statistically associated with a trait or disease of interest. However, traditional linear regression largely ignores non-additive genetic variation, and the univariate SNP-level mapping approach has been shown to be underpowered and challenging to interpret for certain trait architectures. While machine learning (ML) methods such as neural networks are well known to account for complex data structures, these same algorithms have also been criticized as “black box” since they do not naturally carry out statistical hypothesis testing like classic linear models. This limitation has prevented ML approaches from being used for association mapping tasks in GWA applications. In this talk, we present flexible and scalable classes of Bayesian models which allow researchers to perform simultaneous multi-scale inference on complex traits. While analyzing real data assayed in diverse self-identified human ancestries from the UK Biobank, the Biobank Japan, and the PAGE consortium we demonstrate that interpretable ML has the power to increase the return on investment in multi-ancestry biobanks. Furthermore, we highlight that by prioritizing biological mechanism we can identify associations that are robust across ancestries---suggesting that ML can play a key role in making personalized medicine a reality for all.

Eminent Scholar Talk
Tae Hyun Hwang, Ph.D.
Wednesday, July 19, 2023
9:30 AM – 9:50 AM
St. Petersburg II, III

Dr. Tae Hyun Hwang is an Endowed Cancer Chair at Mayo Clinic Florida, where he leads the Artificial Intelligence (AI) oncology research program. His research focuses on advancing the understanding of complex biological systems through the innovative integration of AI, machine learning, and computational methods with experimental approaches. This interdisciplinary work covers spatial biology, single-cell biology, digital pathology, biomarker discovery, and therapeutic development, including immunotherapy and cellular therapy strategies. A significant aspect of Dr. Hwang's research involves the development and analysis of 3D tumor atlas models at subcellular-resolution. These models provide valuable insights into the tumor immune microenvironment (TIME) and its complex role in various cancer types. By deciphering the relationships between cellular and molecular components, his comprehensive approach guides the identification of novel biomarkers and informs the development of personalized therapeutic interventions, including cutting-edge immunotherapies and cellular therapies such as CAR-T/NK cell therapy. Utilizing the power of AI, machine learning, and computational methods in harmony with robust experimental approaches, Dr. Hwang's research is at the forefront of enhancing our understanding of multifaceted biological systems. His ultimate objective is to revolutionize cancer treatment by formulating personalized therapies that enhance patient outcomes.

Title: Harnessing single cell and spatial genomics with machine learning and AI to develop biomarker and therapeutic strategies for immuno and cellular therapy in cancer

In this talk, I will present how single-cell and spatial genomics, combined with machine learning and AI, can revolutionize biomarker identification and the development of personalized immunotherapy and cellular therapy strategy for cancer treatment. By analyzing sequential single-cell CAR-T and PBMC data from CD19 CAR-T cell treated patients, we aim to uncover the dynamic interplay between circulating CAR-T cells, immune suppressive cells, and the immune system in the context of CAR-T cell therapy. Additionally, I will discuss AI-guided biomarker discovery and therapeutic strategies specifically tailored for gastric cancer immunotherapy, which are paving the way for more targeted and effective treatments in the rapidly evolving field of oncology. Lastly, I will present our ongoing efforts to integrate holotomography with spatial transcriptome, with the aim of unveiling novel insights into cellular communication at the subcellular level.

Technology Session
Sunday, July 16, 2023
2:00 PM – 4:45 PM
St. Petersburg I

Chair: Zhongming Zhao

Speaker: Dr. Ryan Mote (10x Genomics)
Title: Unveiling the complexity of breast cancer through advanced analysis of FFPE tissue: single cell, spatial, and in situ mapping of the tumor microenvironment
Abstract: The vast complexities of cancer are characterized by heterogeneity across samples, from tumor cells and tumor microenvironments to therapeutic responses. Elucidating disease mechanisms requires a deep understanding of these complexities. However, traditional assays and tools that analyze tissue in bulk miss significant amounts of information and context due to limited throughput and/or resolution. Through innovations in single cell sequencing and spatial transcriptomics, solutions from 10x Genomics help researchers investigate the body's response to tumors, discover tumor-associated mutations, and uncover mechanisms of acquired resistance to therapy. Join 10x Genomics to learn how researchers are using our single cell assays and spatial tools to gain a multidimensional view of cancer.
Keywords: 10x Genomics, Single Cell Sequencing, spatial transcriptomics, in situ transcriptomics, FFPE

Speaker: Dr. Shan Yang (Complete Genomics)
Title: StereoCell: A bioinformatics tool enables accurate single-cell segmentation for spatial transcriptomics dataset
Abstract: Recent advances in resolution and field-of-view (FOV) have enabled spatially resolved omics to emerge as cutting-edge technologies that provide a technical foundation for interpreting biological signals in large tissue areas at the single-cell level. However, it is challenging to handle the high-resolution spatial omics dataset with associated images and generate spatial single-cell level expression. Here, we propose StereoCell, an image-facilitated cell segmentation framework for high-resolution and large FOV spatial omics. StereoCell offers a comprehensive and systematic solution for generating high-confidence spatial single-cell expression profiles, including image stitching, registration, nuclei segmentation, and molecule labeling. In image stitching and molecule labeling, StereoCell delivers the best-performing algorithms by reducing stitching errors and improving the signal-to-noise ratio of single-cell gene expression compared with existing methods. Using mouse brain data, we demonstrated StereoCell's capability of obtaining high-accuracy spatial single-cell expressions that facilitates downstream clustering and annotation.
Keywords: Cell segmentation, spatial omics, image-facilitated, molecule labeling

Speaker: Dr. Yaping Feng (Admera Health)
Title: A scRNA-seq cell type identifying method based on human curated cell marker database and empirical knowledge
Abstract: Identifying the gene markers and the corresponding cell types is an important task in single-cell RNA sequencing (scRNA-seq) analysis. These techniques allow researchers to identify and characterize different cell types within a heterogeneous cell population based on their gene expression profiles.

Unsupervised cell clustering and cell markers derived from cellranger analysis can indicate the cell types for some typical cells. However, many cell types do not have clear and exceptional cell markers. By leveraging the human curated cell markers' database CellMarker2.0, we developed the rank and fold change scoring algorithm (RFCSA) to connect the cell clusters with the cell types from the database and the researcher' empirical evidence. The scoring algorithm RFCSA was successfully applied to identifying the cell types of these tissues with or without diseases: blood, breast, liver, lung, kidney, knee, brain, and adipose.
Keywords: Single cell, gene marker, cell clustering, cell type identification

Speaker: Dr. Yi Guo (University of Florida)
Title: Generating real-world evidence using OneFlorida+ clinical research consortium
Abstract: The rapid adoption of electronic health record (EHR) systems in the past decade has made large collections of longitudinal and detailed clinical data available for research. The US Food and Drug Administration (FDA) uses the term real-world data (RWD) to refer to information derived from sources outside research settings, including EHRs, administrative claims, and billing data among others. The FDA has recently approved the use of RWD such as those in EHRs to provide real-world evidence in support of the effectiveness or safety for a new drug approval or labeling changes for an approved drug. The importance of RWD is also underscored by the national Patient-Centered Clinical Research Network (PCORnet), a large and highly representative infrastructure containing RWD for over 80 million patients, funded by the Patient-Centered Outcomes Research Institute (PCORI). As one of the eight networks in the national PCORnet, the OneFlorida+ Clinical Research Consortium manages a centralized research patient data repository (OneFlorida+ Data Trust) that contains robust, linked longitudinal patient-level RWD for 16.8 million patients in Florida, 2.1 million patients in Georgia (via Emory), and 9.8 thousand patients in Alabama (via UAB Medicine). In Florida, ten major healthcare systems contribute EHR data to the Data Trust, covering the major metropolitan regions in Florida including Miami, Orlando, Tampa, Jacksonville, Tallahassee, and Gainesville. Research usage of the Data Trust has been significant, including major projects funded by PCORI and the National Institutes of Health.
Keywords: Data Warehousing; Medical Records; Data Analysis; Translational Research

Speaker: Drs. Min Zhang & Bi Zhao (University of South Florida)
Title: Unraveling the Challenges of Genomic Sequencing and Computational Analysis: Introducing the Genomics Sequencing Core and Computational Core
Abstract: Genomic data analysis plays a crucial role in genetic and genomic studies, enabling scientists and researchers to extract valuable insights from vast amounts of genomic data. However, the journey from sample preparation to genomic sequencing and data analysis presents numerous challenges that can significantly impact the accuracy and reliability of results. This presentation aims to explore common troubleshooting scenarios encountered throughout the sequencing process, including sample quality control, library construction, and addressing sequencing errors. By examining these challenges, we seek to equip researchers with practical insights and troubleshooting techniques to overcome issues in sample preparation, sequencing, and subsequent computational steps. The integration of the USF Genomics Sequencing Core and Computational Core services offers a comprehensive solution to tackle these challenges effectively. With practical examples, we will delve into the strategies employed by these cores to address common issues and ensure accurate and reliable data analysis. Join us as we navigate the intricacies of genomics research, gain valuable insights into effective troubleshooting strategies, and discover the capabilities of the USF Genomics Sequencing Core and Computational Core. The USF Genomics Core Facilities, together

with the USF Omics Hub, provide sequencing and computational consulting, data analysis of various sequencing types, database management, and pipeline design and development. With user-friendly pipelines available through the USF Omics Hub on GitHub, researchers can efficiently harness the power of genomics and maximize the potential of their data.

Workshop – Applications of AI in Translational Research
Sunday, July 16, 2023
2:00 PM – 4:45 PM
St. Petersburg II, III

Chairs: Zhe He, Rui Yin

Translational research is the process of transforming scientific discoveries into practical applications and artificial intelligence (AI) has the potential to revolutionize translational research by enabling faster and more accurate data analysis, predicting drug efficacy, and identifying new targets for drug development. This workshop will focus on the growing applications of AI in translational research, which involves the translation of scientific discoveries into practical applications, such as AI-based drug discovery and development, clinical trial optimization and new treatments or therapies for patients. The attendees will discuss the opportunities and challenges of using AI in translational research and explored ways to overcome the limitations and ethical concerns. We will have experts from academia, industry, and government agencies present their latest research and share their experiences in the applications of using AI in translational research. This workshop will provide an opportunity for participants to discuss the challenges and limitations of using AI in translational research, including ethical considerations, data privacy, and regulatory issues. Overall, it will also provide a platform for participants to exchange ideas, share experiences, and identify new opportunities for using AI in translational research.

Speaker: Dr. Qianqian Song (Wake Forest University)
Title: Adaptive graph model deciphers spatial cellular communications
Abstract: Cell–cell communications are vital for biological signaling and play important roles in complex diseases. Recent advances in single cell spatial transcriptomics (SCST) technologies allow examining the spatial cell communication landscapes and hold the promise for disentangling the complex ligand–receptor (L–R) interactions across cells. However, due to frequent dropout events and noisy signals in SCST data, it is challenging and lack of effective and tailored methods to accurately infer cellular communications. To address these challenges, we have proposed a novel adaptive graph model with attention mechanisms named spaCI. spaCI incorporates both spatial locations and gene expression profiles of cells to identify the active L–R signaling axis across neighboring cells. Through benchmarking with currently available methods, spaCI shows superior performance on both simulation data and real SCST datasets. spaCI achieves to reveal hidden L–R interactions and their upstream transcription factors from different types of SCST data such as seqFISH+ and NanoString CosMx Spatial Molecular Imager (SMI) data. Collectively, spaCI addresses the challenges in interrogating SCST data for gaining insights into the underlying cellular communications, thus facilitates the discoveries of disease mechanisms, effective biomarkers and therapeutic targets.

Speaker: Dr. Rui Yin (University of Florida)
Title: Leveraging the power of genomics to facilitate the diagnosis of undiagnosed diseases with machine learning models
Abstract: Rare and ultra-rare genetic conditions are estimated to impact nearly 1 in 20 people worldwide, yet accurately pinpointing the diagnostic variants underlying each of these conditions remains a formidable challenge. Because comprehensive, in vivo functional assessment of all possible genetic variants is infeasible, clinicians instead consider in silico variant pathogenicity predictions when interpreting variants of uncertain significance. In the most difficult undiagnosed cases, such as those accepted to and profiled in the Undiagnosed Diseases Network (UDN), existing pathogenicity predictions fail to score or tend to mischaracterize patients' disease-causing genetic variants. Here, we presented VarPPUD, a random forest-based variant pathogenicity predictor trained specifically on variants implicated in UDN cases that considers gene-, amino acid- and nucleotide-level features. VarPPUD achieves a cross-validated accuracy of 79.3% and precision of 77.5% on a held-out subset of these uniquely challenging cases, respectively representing an average 18.6% and 23.4% improvement over nine related approaches. We further validate VarPPUD's discriminatory ability on GAN-generated synthetic variants as well. Finally, we demonstrate how our model is amenable to evaluating each input feature's importance and contribution toward prediction—an essential step toward understanding underlying mechanisms of newly-uncovered disease-causing variants.

Speaker: Dr. Mei Liu (University of Florida)
Title: Better Acute Kidney Injury Prediction and Risk Factor Analysis with Personalized Transfer Learning
Abstract: Acute kidney injury (AKI) is a life-threatening clinical syndrome characterized by rapid reduction of kidney function and has complex etiologies and pathogenesis. Prevalence of hospital-acquired AKI varies by patient population, affecting 7% to 18% of general inpatients and greater than 50% of patients in the intensive care unit. Complex risk factors and their interactions hinder physicians from forecasting AKI risk. I will present one of our recent published work on the development and validation of a personalized AKI risk prediction model using electronic health records (EHR). Results from our study demonstrated that a personalized modeling with transfer learning is an improved AKI risk estimation approach that can be used across diverse patient subgroups. Risk factor heterogeneity and interactions discovered at the individual level highlighted the need for agile, personalized care.

Speaker: Dr. Zhe He (Florida State University)
Title: Harnessing Explainable, Equitable, and Actionable AI to Improve Health
Abstract: Predictive modeling of health outcomes using artificial intelligence (AI) has recently revolutionized healthcare. The availability of large amounts of data (e.g., electronic health records (EHRs)) along with a significant increase in computational power has enabled researchers to further investigate the benefits of applying AI to predictive analytics for medicine and healthcare. While prior research has shown superior performance of deep learning when predicting health outcomes using electronic health record (EHR) data, it has not been adequately adopted in EHR systems in the US. Evidence has shown that improved transparency and interpretability of the deep learning models will increase their trustworthiness for medical doctors, thereby increasing their adoption by healthcare systems. In this talk, I will discuss our recent research efforts on enhancing the interpretability of machine learning and deep learning models for predicting health outcomes among patients with cardiovascular diseases using EHR data. This research aims to improve the use of AI-based decision support systems by optimizing the balance between model performance and interpretability.

Speaker: Dr. Jinfeng Zhang (Florida State University)
Title: Constructing a Large-Scale Biomedical Knowledge Graph and Its Applications in Drug Discovery
Abstract: In the past few decades, the biomedical research community has acquired a wealth of knowledge, much of which is stored in scientific literature as unstructured text. Converting this text into structured form is crucial for developing new methodologies and applications that can fully utilize this knowledge. To achieve this goal, two basic problems must be addressed: named entity recognition (NER) and relation extraction (RE). NER involves identifying the concepts or entities in texts, such as diseases, genes/proteins, and chemical compounds. RE, on the other hand, aims to extract the relationships between these entities. The information extracted from NER and RE can be used to create knowledge graphs, where nodes represent entities in the text and edges represent their relationships. This presentation will discuss our team's work on the LitCoin NLP Challenge organized by NIH, for which we were awarded first place. Using pipelines developed for the challenge, we processed all PubMed articles and created a large-scale biomedical knowledge graph. The accuracy of this large-scale relation extraction is estimated to be 84% based on manual verification of a sample of the extracted data. We also incorporated relation information from 40 public databases and relations inferred from publicly available genomics datasets. Our knowledge graph consists of over 11 million entities and more than 40 million relations. We have developed versatile query functions and knowledge discovery tools for accessing and mining structured data in the knowledge graph. Finally, we will discuss some drug discovery-related applications enabled by this large-scale knowledge graph.

Speaker: Dr. Lai Wei(Ohio State University)
Title: Translational Pharmaco-informatics research
Abstract: In this talk, I am going to cover three translational pharmaco-informatics research topics. In the first topic, we will introduce how to generate pharmacogenetic hypotheses from drug interaction evidence. In the second topic, we will demonstrate a knowledge base, called Drug Combo. We will illustrate how it can assist Phase I drug combination trial design. In the last topic, we will present a knowledgebase, MPRINT-KB. We will introduce how this knowledgebase can guide maternal and pediatric precision therapeutics research, with the ultimate goal of changing drug labels for maternal and pediatric patient populations.

Speaker: Dr. Yi Luo (Moffitt Cancer Center)
Title: Using Explainable Machine Learning Models to Predict CAR T-Cell Therapy Response with Longitudinal Patient Report Outcomes
Abstract: Introduction: More than 50% of patients with advanced hematologic cancers have limited remaining treatment options. Chimeric antigen receptor T (CAR T) -cell therapy is a revolutionary treatment to them that results in durable remission and / or cure. However, most recipients developed unique toxicities that can be life threatening if not identified and treated early. As feedback from recipients before and during the CAR T-cell treatment, patient-reported outcomes (PROs) have potential of predicting important clinical outcomes. The participants' PRO Measurement Information System (PROMIS) includes their anxiety, depression, fatigue, sleep disturbance, physical function, and pain interference information. In order to develop appropriate intervention / care to improve the outcomes and allow clinicians to identify patients who are most suitable for inpatient or outpatient CAR T-cell treatment, we intended to develop risk prediction models for CAR T-cell toxicities and outcomes before and during the therapy. As a primary study to achieve this goal, we would like to predict CAR T-cell therapy response based on the patients'

demographic, clinical characteristics, and PROs in this study. **Method**: 208 participants were collected from four studies with common PROs at pre-CAR T-cell infusion (baseline) and on 90 days post-infusion. Only 62 of them had Day 30 PROs. Patients with partial response or better and patients with stable / progressive disease or death were grouped into responders and non-responders, respectively. Three Bayesian network (BN) models were developed to predict 90-day CAR T-cell therapy response. Model A included baseline patient characteristics and PROs. Model B added change in each PRO from baseline to day 90. Model C added change in PROs from baseline to day 30. The predictive performance of these models was evaluated with area under the receiver operating characteristics curve (AU-ROC) and 95% confidence intervals (CI) based on 2,000 stratified bootstrap replicates. **Results**: Model C had the best predictive performance (AU-ROC = 0.83, 95% CI: 0.72 - 0.91) and significantly outperformed Models A and B based on DeLong's test. In model C, there were direct arcs to the response from multiple myeloma diagnosis, diffuse large B-cell lymphoma diagnosis, baseline physical and cognitive functions, 30-day increased physical function, and 90-day increased depression and anxiety. **Conclusions**: BN models identified important demographic, clinical features, and PROs related to CAR T-cell therapy response before and during the treatment and displayed transparent pathways to predict and intervene the response. However, they still need to be tested in independent and external datasets.

Workshop – Microbiome Data Analysis
Monday, July 17, 2023
9:30 AM – 12:00 PM
St. Petersburg II, III

Chair: Qunfeng Dong

Speaker: Dr. Noam Auslaner (The Wistar Institute)
Title: Artificial intelligence-based identification of microbes in cancer
Abstract: Shifts in tumor microbiomes are increasingly recognized as a key factor modulating the development and progression of cancers. Short-read RNA sequencing is perhaps the most widely used techniques in life science, including cancer research. Yet, study of the microbial components expressed in tumors from RNA sequencing is currently challenging. We therefore develop artificial intelligence-based methods for virus and bacteria identification, to establish efficient tools that can detect microbial expression in human tissues through RNA sequencing. Our new approach allows characterization of expressed microbial species and proteins in human cancers and healthy tissues through RNA sequencing. Applying our method to study different cancer types, we detect new and divergent viruses that have not been reported in cancers before. We identify several bacterial clades whose prevalence is linked to cancer development. In addition, we uncover bacterial protein families that are significantly associated with patient survival, providing additional axes for cancer subclassification and prognosis. Overall, we reveal different microbes that influences cancer development, and provide a framework for the analysis of other human malignancies whose development may be driven by pathogens.

Speaker: Dr. Ying Zhang (University of Rhode Island)
Title: Metagenomic read classification using deep learning models

Abstract: Metagenomics is a technique for genome-wide profiling of microbiomes. Broad applications of this technology have led to the accumulation of short DNA sequences, namely metagenomic reads, from diverse microbiomes. Identification of species and functional composition requires advanced computational tools to enable a more efficient workflow for analyzing such massive data. DL-TODA is a deep learning based approach for the rapid classification of metagenomic reads into over 3000 bacterial species. It applies a convolutional neural network architecture for the modeling of species-specific features. In this presentation, we will compare DL-TODA with other existing metagenome read classification tools and demonstrate its application in microbiome research.

Speaker: Dr. Yunxi Liu (Rice University)
Title: Squeegee: De Novo Identification of Contaminants in Low Microbial Biomass Microbiomes
Abstract: Computational analysis of host-associated microbiomes has opened the door to numerous discoveries relevant to human health and disease. However, contaminant sequences in metagenomic samples can potentially impact the interpretation of findings reported in microbiome studies. Negative control experiments are considered the standard for contamination removal; however, negative control data are commonly unavailable in public databases. Here we present Squeegee, a de novo contamination detection tool that facilitates the detection of microbial contaminants when negative controls are not available. Squeegee is based on the hypothesis that microbial contaminants from the same source, such as DNA extraction kits or from a lab environment, will share similar characteristics in the composition of their contaminants. To test this hypothesis, we compared Squeegee microbial contaminant predictions to experimental negative control data and show that Squeegee accurately recovers putative contaminants. After analyzing samples of varying biomass from the Human Microbiome Project, our method identifies likely, previously unreported kit contamination. To further evaluate Squeegee, we benchmarked it against a gold-standard contamination detection approach in Decontam. We found Squeegee can achieve performance that meets or exceeds Decontam predictions at species rank in some cases, with respect to unweighted F-score, F-score weighted by the relative abundance in the non-control samples, and cumulative relative abundance of the putative correctly identified contaminants from the negative control experiment samples. Furthermore, Squeegee achieves high weighted recall (weighted by both relative abundance of taxa in negative control and non-control samples) and low false positive rates on real metagenomic datasets. In summary, Squeegee can help to identify putative contaminant sequences of suspicious taxa for low-biomass microbiome studies, enabling sample-independent and orthogonal approaches aimed at distinguishing true microbiome signals from environmental contamination. Squeegee is open-source and available for download at https://gitlab.com/treangenlab/squeegee.

Speaker: Dr. Shulei Wang (University of Illinois Urbana-Champaign)
Title: Multiscale adaptive differential abundance analysis in microbial compositional data
Abstract: Differential abundance analysis is an essential and commonly used tool to characterize the difference between microbial communities. However, identifying differentially abundant microbes remains a challenging problem because the observed microbiome data is inherently compositional, excessive sparse, and distorted by experimental bias. Besides these major challenges, the results of differential abundance analysis also depend largely on the choice of analysis unit, adding another practical complexity to this already complicated problem. In this work, we introduce a new differential abundance test called the MsRDB test, which embeds the sequences into a metric space and integrates a multi-scale adaptive strategy for utilizing spatial structure to identify differentially abundant microbes. Compared with existing methods, the MsRDB test can detect differentially abundant microbes at the finest resolution offered by data and

provide adequate detection power while being robust to zero counts, compositional effect, and experimental bias in the microbial compositional data set.

Speaker: Dr. Huilin Li (New York University)
Title: A novel microbial causal mediation analytic pipeline for investigating microbiome's mediating role in disease and health disparity
Abstract: An important hallmark of human microbiota is its modifiability and dynamics. Many microbiome association studies have revealed the important association between microbiome and disease/health status, which encourage people to dive deeper to uncover the causation of microbiota in the underlying biological mechanism. This enthusiasm has opened the way to apply statistical causal models to quantify the causal effect of microbiota and to identify the specific acting microbes in the study. Here, we propose a rigorous Sparse Microbiome Causal Mediation Model (SparseMCMM) specifically designed for the high dimensional and compositional microbiome data. In particular, the Dirichlet multivariate regression model and linear log-contrast regression model are proposed to estimate the causal direct effect of treatment and mediation effect of the microbiota at both community level and individual level. Regularization techniques are used to perform the variable selection in the proposed model framework to identify signature causal microbes. We exhibit the utility of SparseMCMM in both clinical trial and observational study design through real data analyses.

Speaker: Mahsa Monshizadeh (Indiana University)
Title: Incorporating metabolic activity, taxonomy and community structure to improve microbiome-based predictive models for host phenotype prediction
Abstract: The human gut microbiome plays key roles in human health and diseases. We developed MicroKPNN, a prior-knowledge guided interpretable neural network for microbiome-based human host phenotype prediction. The prior-knowledge used in MicroKPNN includes the metabolic activities of different bacterial species, phylogenetic relationships, and bacterial community structure. Application of MicroKPNN to seven gut microbiome datasets (involving five different human diseases including inflammatory bowel disease, type 2 diabetes, liver cirrhosis, colorectal cancer, and obesity) shows that incorporation of the prior knowledge helped improve the microbiome-based host phenotype prediction. MicroKPNN outperformed fully-connected neural network based approaches in all seven cases, with the most improvement of accuracy in the prediction of type 2 diabetes. MicroKPNN outperformed a recently developed deeplearning based approach DeepMicro, which selects the best combination of autoencoder and machine learning approach to make predictions, in six out of the seven cases. More importantly, we showed that MicroKPNN provides a way for interpretation of the predictive models. Our results suggested that the metabolic potential of the bacterial species contributed more than the two other sources of prior knowledge. MicroKPNN is publicly available at https://github.com/mgtools/MicroKPNN.

Speaker: Xiang Gao (Loyola University Chicago)
Title: Exploring the Male Urethral Microbiome: A Community Ecology Approach Based on the Neutral Theory
Abstract: The Unified Neutral Theory of Biodiversity and Biogeography, which proposes that ecosystem diversity and composition largely result from random processes such as migration, birth, death, and speciation, rather than deterministic processes, offers an important framework for studying the human microbiome. This study applies the Neutral Theory as a simple null model to explore the composition of the human male urethral microbiome, which harbors a relatively simple core microbial community. If the

actual composition of the microbiome significantly deviates from predictions made by the Neutral Theory, it would suggest a substantial role for deterministic processes like selection. Therefore, our investigation aims to discern whether the male urethral microbial species coexist due to such random processes, rendering them "neutral" with respect to each other, or if they are subject to positive or negative selection. Our findings also delve into the role of migration in urethral microbiome diversity. This aligns with our recent discovery of a strong correlation between sexual behaviors and the inter-specimen variance in urethral microbiome composition. Overall, this study contributes to understanding the dynamic and complex nature of the human male urethral microbiome.

Workshop – Prompt Bioinformatics: Application of ChatGPT and Large Language Models
Monday, July 17, 2023
1:40-4:15 PM
St. Petersburg I

Chair: Gangqing Hu

Labs in the pBio group: Mei Chen (Microsoft); Xijin Ge (South Dakota State University); Wenpin Hou (PI), Wenhan Bao (Columbia University); Gangqing Hu (PI), Li Ma, Weijun Yi (West Virginia University); Zhicheng Ji (Duke University); Li Liu (Arizona State University); Tao Liu (PI), Zhou Jiaojiao (Roswell Park Comprehensive Cancer Center); Sayaka Miura (Temple University); Zhengwei Xie (Peking University); Dong Xu (PI), Yibo Chen (University of Missouri); Pingkun Yan (PI), Xuanang Xu, Jiajin Zhang (Rensselaer Polytechnic Institute); Qiuming Yao (PI), Pengchong Ma (University of Nebraska - Lincoln); S. Stephen Yi (PI), Xingxin Pan (The University of Texas at Austin); Erliang Zeng (PI), Phuong Fawng Nguyen, Shri Vishalini Rajaram (University of Iowa); Chan Zhou (PI), Zixiu Li, Peng Zhou (University of Massachusetts Chan Medical School); Fengfeng Zhou (PI), Yusi Fan, Kewei Li (Jilin University); Wanding Zhou (PI), Hongxiang Fu (University of Pennsylvania); Xiang Zhou (PI), Katherine Zhou (University of Michigan); Qiyun Zhu (PI), Zhu Xing (Arizona State University).

Speaker: Dr. Gangqing Hu (West Virginia University)
Title: Prompt Bioinformatics with Chatbots
Abstract: We introduce Prompt Bioinformatics, a transformative approach where a chatbot generates code for intricate bioinformatics data analysis based on natural language instructions. With accumulating evidence underscoring the success of ChatGPT in small and discrete bioinformatics tasks, we seek to address a crucial question: How to effectively steer a chatbot for a comprehensive bioinformatics data analysis? The pBio group (see below) is a collaborative initiative, keenly exploring the potential of ChatGPT in complex bioinformatics applications. Developing strategic approaches to mitigate response uncertainties and maximize result reproducibility is one of our key goals. A significant contribution from the group is a detailed knowledgebase, encapsulating a spectrum of case studies related to different areas of bioinformatics data analysis using a prompt-based methodology. In this presentation, I will elucidate a generic framework for building a prompt-based pipeline, using the ATAC-seq case study, adept at managing complex bioinformatics analyses. Further case studies will be presented by other speakers in separate talks throughout this special session.

Speaker: Dr. Sayaka Miura (Temple University)
Title: PROMPT BIOINFO. CASE STUDY: Intra-tumor Evolutionary Inference
Abstract: Tumorigenesis is an evolutionary process in which mutation and selection drive the growth of a single clone into diverse populations of cells (i.e., subclones), leading to intra-tumor heterogeneity. Spatiotemporal changes of subclones have been associated with clinical characteristics of a tumor, such as metastasis and treatment outcomes. Understanding the evolutionary dynamics of a tumor is a critical step toward precision oncology. In this case study, we will present how to use GPT-4 to assist intra-tumor evolutionary inference based on somatic mutations detected via bulk and single-cell DNA sequencing. The tasks include subclone identification, phylogeny inference, and visualization. We show that GTP-4 can generate R and Python scripts to format input files and perform data analysis using available computational methods and tools. We will also discuss tips for prompting to generate correct scripts, as well as limitations.

Speaker: Dr. Chan Zhou (University of Massachusetts)
Title: Leveraging Stand-alone RNA-Seq Data for Novel lncRNA Identification and Annotation: A Prompt Bioinformatics Case Study
Abstract: Long non-coding RNAs (lncRNAs) play pivotal roles in biological processes and diseases, yet many remain undiscovered due to their specific expression in disease states. High-throughput sequencing techniques have provided extensive transcriptomics data, but their potential has been limited by traditional methodologies' constraints. To address this, we developed Flnc, a tool that identifies lncRNAs from standalone RNA-seq data with an 85% prediction accuracy (Li et al., Non-coding RNA 2022), significantly surpassing conventional methods' 50% accuracy rate. Flnc also uniquely identifies single-exon and human-specific lncRNAs. Our presentation will introduce a prompt-based pipeline, utilizing ChatGPT, that employs Flnc for identifying and annotating novel lncRNAs. This approach advances our understanding of lncRNA's roles in disease pathogenesis. To construct this pipeline, we use structured natural language prompts, enabling ChatGPT to generate the required code. This robust pipeline has undergone rigorous validation tests, further ensuring its efficacy.

Speaker: Dr. Li Liu (Arizona State University)
Title: Exploring ChatGPT's Ability to Generate Novel Algorithms in Bioinformatics
Abstract: The application of large language model (LLM)-based chatbots, such as ChatGPT, in generating programming codes to aid bioinformatics analysis has gained attention. However, it is important to consider that existing pipelines and example codes, which are part of the training data for LLMs, are readily available online. This raises questions about the ability of ChatGPT to develop novel algorithms for bioinformatics analysis. To assess its capability in this regard, I conducted an experiment where I tasked ChatGPT with designing a new algorithm to decompose mixed distributions of bulk sequencing data and implementing it in R. ChatGPT proposed several statistical models and utilized an Expectation-Maximization algorithm for implementation. Although the initial code did not yield the expected results, ChatGPT succeeded in producing the desired outcome after receiving more specific instructions. In this presentation, I will share my experience and the valuable lessons learned from this experiment.

Speaker: Yibo Chen (University of Missouri)
Title: Enhanced Gene Interaction Analysis and Pathway Reconstruction through Iterative Prompt Refinement by ChatGPT

Abstract: ChatGPT trained by massive text contains valuable scientific information. It is a great opportunity to mine ChatGPT as a knowledge graph for research purposes. However, it is challenging to mine useful information while simultaneously curbing hallucination output. There is no systematic method to address this challenge as of now. This study explores various prompt techniques to optimize ChatGPT's abilities in a bioinformatics use case and systematically evaluate the results. The use case is to decipher complex gene interactions, such as activation, inhibition, phosphorylation, and ubiquitination from ChatGPT. The ground truth data for the evaluation is from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database, including 200 training samples (80 activations, 80 inhibitions, and 40 phosphorylations) as well as 50 each for validation and testing samples (20 activations, 20 inhibitions, and 10 phosphorylations). Initial assessments of ChatGPT were performed by prompting ChatGPT to provide binary answer to questions on whether two genes have a specific relation. Straightforward gene relation queries yielded a precision of 0.95, recall of 0.67 and F-1 score of 0.333 for the testing samples. The feedback mechanisms subsequently bolstered the response quality. In particular, after reminding ChatGPT that two genes have a specific relation, it is more likely to provide a better answer to the same question. We also reconstructed the non-small cell lung cancer pathway using ChatGPT. Initially, the reconstructed pathway could mine merely 19.4% of genes from the KEGG database. However, using the 'least-to-most' prompting strategy to reconstruct gene pathways can mine 27.8% of KEGG genes. To further enhance ChatGPT's responses, we embarked on an iterative prompt refinement strategy. In each iteration, we selected previous successful prompts assessed in the training samples. Then we fed their prompts as well as assessment metrics (e.g., F1 score, precision, and recall) and suggestions on how to write a better prompt, such as changing the role of prompts or paraphrasing the prompt, into ChatGPT for suggesting better prompts. ChatGPT displayed remarkable capability in providing novel prompts and iteratively enhancing prompts based on implicit feedback from previous iterations. For instance, when tasked with unraveling the relationship between two genes using Tree-of-Thought prompts increased the F-1 score from 0.333 to 0.396. After three iterations, ChatGPT further boosted the F-1 score to 0.538, highlighting its potential as an automated 'prompt engineer' with a performance surpassing manually curated prompts. This approach presents a novel strategy to optimize prompts to mine knowledge from ChatGPT in general.
Keywords: ChatGPT, Prompt Engineering, Prompt Optimization, Bioinformatics, Gene Pathway, KEGG Pathway Database

Speaker: Shouvon Sarker (Prairie View A&M University)
Title: Ensemble BERT for Medication Event Classification on Electronic Health Records (EHRs)
Abstract: Identification of key variables such as medications, diseases, relations from health records and clinical notes has a wide range of applications in the clinical domain. n2c2 2022 provided shared tasks on challenges in natural language processing for clinical data analytics on electronic health records (EHR), where it built a comprehensive annotated clinical data Contextualized Medication Event Dataset (CMED). This study focuses on subtask 2 in Track 1 of this challenge that is to detect and classify medication events from clinical notes through building a novel BERT-based ensemble model. It started with pretraining BERT models on different types of big data such as Wikipedia and MIMIC. Afterwards, these pretrained BERT models were fine-tuned on CMED training data. These fine-tuned BERT models were employed to accomplish medication event classification on CMED testing data with multiple predictions. These multiple predictions generated by these fine-tuned BERT models were integrated to build final prediction with voting strategies. Experimental results demonstrated that BERT-based ensemble models can effectively improve strict Micro-F score by about 5% and strict Macro-F score by about 6%, respectively.

Keywords: Electronic Health Records, Medication Events, Bidirectional Encoder Representations from Transformers (BERT), Ensemble Model

Speaker: Cong Liu (Children's Hospital of Philadelphia)
Title: Enhancing Phenotype Recognition in Clinical Notes Using Large Language Models: PhenoBCBERT and PhenoGPT
Abstract: To understand the phenotypic presentation of genetic diseases, various Natural Language Processing (NLP) methods have been developed to identify Human Phenotype Ontology (HPO) within clinical notes. HPO provides a standardized vocabulary of phenotypic abnormalities, and was predominantly curated by domain-specific experts. However, some clinical phenotypes may not be well represented by the current HPO vocabulary. In addition, existing HPO tagging tools are typically based on sets of heuristics or rules, overlooking contexts of the phenotype descriptions. We hypothesize that large language models (LLMs) based on the transformer architecture can enable automated detection of clinical phenotype terms, including terms not documented in the HPO. In this study, we developed two methods, including PhenoBCBERT, a BERT-based model leveraging Bio+Clinical BERT as the pretrained model, as well as PhenoGPT, a GPT-based model initialized from several latest GPT models (GPT-J, GPT-3, and GPT-3.5). We compared our methods with PhenoTagger, a recently developed HPO recognition tool that combines rule-based and deep learning methods. We found that our methods can extract more phenotype concepts, including novel ones not characterized by HPO. We also performed case studies on biomedical literature to illustrate how new phenotype information can be recognized and extracted. We compared current BERT-based versus GPT-based models for phenotype tagging, in multiple aspects including model architecture, memory usage, speed, accuracy, and privacy protection. We also discussed the addition of a negation step and an HPO normalization layer to the transformer models for improved HPO term tagging. In conclusion, PhenoBCBERT and PhenoGPT enable the automated discovery of phenotype terms from clinical notes and biomedical literature, facilitating automated downstream tasks to derive new biological insights on human diseases.

Tutorial – Collection and Analyses of Long-Read Transcriptome and Epitranscriptome Data
Monday, July 17, 2023
9:30AM – 12:00 PM
St. Petersburg I

Organizer: Kin Fai Au (University of Michigan)

Long-read sequencing techniques developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have revolutionized omics analyses at the single-molecule level and long-range scale, garnering significant attention in the biomedical research community. Long-read sequencing offers unique advantages in various basic research and clinical applications. For example, long-read transcriptome sequencing enables the capture of the full-length gene isoforms/transcripts, providing more accurate transcriptome identification and quantification as well as analyses of complex transcriptional events. Additionally, the ONT direct RNA sequencing technique offers the chance to directly detect RNA modifications, such as N^6-methyladenosine (m^6A) and Adenosine-to-Inosine (A-to-I) editing.

Although long-read sequencing brings unprecedented opportunities for solving numerous biological challenges, researchers and biologists who are not familiar with the technology or data analysis may encounter difficulties in fully utilizing it to boost their research, due to several major barriers:

(1) Lack of best-practice guidelines for designing the sequencing strategies, conducting experiments, generating high-quality data, and optimizing the analysis frameworks;
(2) Lack of computational resources to support the storage, management and analysis of the long-read data;
(3) Lack of the datasets from various biological samples and tissues, which limits the broader adoption of long-read sequencing in different research areas.

To bridge the computational gap, we have developed a user-friendly long-read data analysis web platform which currently supports the analysis of the long-read transcriptome and epitranscriptome data. Supported by the evaluation results from Long-read RNA-seq Genome Annotation Assessment Project (LRGASP) Consortium (https://www.gencodegenes.org/pages/LRGASP/), the platform implements the best-practice pipeline, integrates the widely-used long-read bioinformatics analysis tools, and provides a comprehensive visualization panel to gain an in-depth understanding of the analysis results.

In this tutorial, we will firstly overview the long-read sequencing techniques, introduce the experimental design and data collection specific to long-read transcriptome sequencing, and offer a step-by-step analysis guideline and hands-on experience in analyzing long-read transcriptome data using the platform. Our tutorial will also cover the latest benchmarking and evaluation resources and results for long-read-based transcriptomics analysis.

Learning Objectives:

By the end of this tutorial, participants will have achieved the following learning objectives:

- Understand the principles and advancements of PacBio and ONT long-read sequencing technologies;
- Stay updated with the latest developments in the field of long-read sequencing;
- Gain knowledge of different experimental designs suited for various research goals;
- Learn best practices for long-read transcriptome analysis, including gene isoform identification and quantification, supported by our latest systematic evaluation work;
- Learn how to detect and visualize the RNA A-to-I editing and m6A detection using ONT direct RNA sequencing data;
- Effectively utilize the web platform to explore, analyze, and derive biological insights from long-read transcriptome data.
- Communicate the findings and biological insights obtained from long-read transcriptome analysis using the web platform.

Intended Audience:

This tutorial is aimed at participants interested in gaining proficiency in long-read sequencing technology, and/or conducting best-practice analysis of long-read transcriptome and epitranscriptome data. It will be particularly valuable for attendees who wish to analyze and utilize the long-read transcriptome data using established best practices, as well as those who may have limited hardware resources to handle such data.

Tutorial – Unlocking the Power of Single-Cell Gene Expression Analysis with SCGEATOOL: A Hands-On Tutorial for Non-Programmers and Machine Learning Experts
Wednesday, July 19, 2023
11:20AM-12:20 PM
St. Petersburg I

Organizer: James Cai (Texas A&M University)

SCGEATOOL (https://scgeatool.github.io/) is an easy-to-use desktop application developed by the workshop instructor for single-cell transcriptome analysis without programming. SCGEATOOL is also an innovative programming ecosystem for data scientists and machine learning experts to deploy advanced analytical functions. This workshop will introduce participants to SCGEATOOL. The application, which can be easily installed on a personal computer, offers a wide range of functionalities for data pre-processing, visualization, and statistical analysis. With simple clicks, participants can explore data with user-friendly interfaces and access updated functions for all essential tasks in single-cell transcriptome analysis. SCGEATOOL supports a wide range of cross-study analyses and data integration to deepen biological understanding. Public data sets in the GEO database can be imported without programming knowledge. Participants will also learn how to import and format their own data, perform basic quality control checks, and perform key analyses such as dimensionality reduction, clustering, and differential expression analysis. SCGEATOOL is optimized to visualize hundreds of thousands of single cells with MATLAB's high-quality interactive techniques, helping users to transform raw in-house data into insights. The workshop will also cover how to combine functions to create novel analysis pipelines. By the end of the workshop, non-programer participants will have the skills to use this application to analyze their own single-cell gene expression data, and machine learning experts will know how this expandable programming ecosystem can help their research. Overall, the workshop aims to give all-level participants hands-on experience and the necessary skills for advanced single-cell data analysis.

Intended Audience:
The target audience for this workshop is researchers and scientists working in the field of molecular biology and genetics, who are interested in analyzing single-cell transcriptome data but have limited programming skills. The workshop would be also suitable for bioinformaticians and data scientists, who are looking for user-friendly tools and interested in adopting new programming ecosystems to develop functions for single-cell biologists.

Rational
Single-cell gene expression analysis is an important and rapidly growing field in molecular biology and genetics, which allows researchers to study the gene expression patterns of individual cells. However, the analysis of this type of data can be complex and requires a significant amount of computational and bioinformatics expertise. This can be a barrier for researchers with limited computational backgrounds, making it difficult for them to access the full potential of their data. The proposed workshop will provide researchers with an opportunity to learn how to analyze their single-cell gene expression data in a user-

friendly way using SCGEATOOL, without needing to have extensive programming skills. SCGEATOOL is a lightweight and blazing-fast desktop application that provides interactive visualization functionality to analyze single-cell transcriptomic data. SCGEATOOL allows participants to easily interrogate different views of their data sets to quickly gain insights into the underlying biology. This workshop will be interactive in nature, allowing participants to work with their own data and see the results of their analysis in real time. This interactive format will be beneficial for participants as it will provide them with the opportunity to ask questions and receive immediate feedback, which will help them to better understand the concepts and methods being presented. Additionally, by working with real data and learning how to analyze it using SCGEATOOL, participants will be able to immediately apply the skills they learn to their own research projects. The hands-on experience will be beneficial for participants as it will give them the necessary skills to analyze their own data and make more informed decisions in their research. Overall, this workshop aims to provide researchers with an accessible, easy-to-use tool for single-cell gene expression analysis, making it possible for researchers with limited computational backgrounds to take full advantage of the insights available from this technology.

Learning objectives

- Participants will be able to install and navigate SCGEATOOL, a desktop application designed to provide interactive visualization functionality to analyze single-cell transcriptomic data and understand the basic features of the software.
- By the end of the workshop, participants will be able to confidently import, format, and perform basic quality control checks on their own single-cell gene expression data using SCGEATOOL.
- Participants will be able to perform key analyses such as dimensionality reduction, clustering, and differential expression analysis on their single-cell gene expression data using SCGEATOOL and understand the basic concepts behind each analysis.
- Expert participants will be able to develop single-cell gene expression data analysis methods that can be incorporated into SCGEATOOL to reach out to broader non-programmer users.

Special Session – Dynamics of Transcriptional Regulation Towards Single Cell, Single Molecular, and Spatial Omics
Tuesday, July 18, 2023
9:30-11:45 AM
St. Petersburg I

Chairs: Kaifu Chen

Speaker: Dr. Chan Zhou (University of Massachusetts)
Title: Uncovering Disease-Associated Novel lncRNAs: A Computational Perspective
Abstract: Long noncoding RNAs (lncRNAs), comprising over 70% of the human genome, remain largely mysterious, even though they play significant regulatory roles in various biological processes and diseases. These RNAs exhibit unique disease-specific expression patterns, underlining their potential as drivers and modifiers of diseases. Yet, due to their often tissue- or cell-type-specific expression in disease states, many lncRNAs remain undiscovered and unannotated. The advent of high-throughput sequencing techniques has unlocked access to transcriptomics data from diverse human biospecimens. We have developed a

computational pipeline that analyzes RNA-seq and matched ChIP-seq data to identify novel lncRNAs. This method has led to the identification and functional prediction of novel lncRNAs, potentially influencing liver fibrosis (Zhou et al., Genome Med. 2016). The predicted functions were later experimentally confirmed in a liver fibrosis disease model (manuscript submitted). To address the common lack of matching ChIP-seq data for most public RNA-seq data, we recently developed Flnc, a tool that identifies lncRNAs from stand-alone RNA-seq data with an impressive 85% prediction accuracy (Li et al., Non-coding RNA 2022). Flnc not only outperforms the conventional method, which has a 50% accuracy rate, but also effectively detects single-exon and human-specific lncRNAs often overlooked by other methods. This advancement enhances our understanding of lncRNA involvement in disease pathogenesis. Flnc is publicly accessible at https://github.com/CZhouLab/Flnc.

Speaker: Dr. Chi Zhang (Indiana University)
Title: Data-driven and AI-empowered systems biology
Abstract: The functional activities of biological systems include both intracellular functions such as transcriptional regulation, metabolism, and signaling transduction, as well as intercellular activities such as cell-cell interactions. In recent years, we focused on developing new systems biology approaches to quantify biological functions and predict biological relations in disease using omics data. One such approach is the single-cell Flux Estimation Analysis (scFEA), a computational method that enables the estimation of cell-wise metabolic flux rates by using single-cell RNA-seq data, and the prediction of how changes in genes and metabolites affect metabolism through in-silico perturbations. Building on scFEA, we have developed a new research framework, "data-driven and AI-empowered systems biology", which aims to quantify biological processes and approximate their dynamic property using omics data. We generalized the analysis from the metabolic system to biosynthesis and processing of large molecules, transcriptional regulation, signaling transduction, and cell-tissue interactions, as well as enabled the usage of multi-omics data. By applying this approach to different disease systems, we have predicted and experimentally validated a range of discoveries, including (1) identifying new drug targets to improve the efficacy of immunotherapy for cancer treatment, (2) predicting the trend of metabolic shifts throughout cancer progression.

Speaker: Dr. Guangyu Wang (Houston Methodist Research Institute)
Title: Deep learning reveals cellular state transition
Abstract: We present a deep learning model, cellDancer, to infer temporal dynamics of cell state transition from the static snapshot of cells based on the RNA velocity which tracks and compares the dynamics of nascent, unspliced mRNAs and relates them to mature, spliced mRNAs. Specifically, cellDancer utilizes single-cell RNA-seq data in a continuous biological process such as cell differentiation and cell cycle to infer 1) the trajectory and pseudotime of cell state transition, 2) the RNA velocity of each cell, and 3) transcription, splicing, and degradation rates of each gene in each cell. Our benchmark datasets showed that cellDancer accurately and robustly infers cell fate transitions in heterogeneous cell populations such as gastrulation erythroid maturation, mouse hippocampus development, human embryonic glutamatergic neurogenesis, and endocrine development. cellDancer also shows an ultra-performance on simulation benchmarks including high dropout ratio datasets and high noise-to-signal ratio datasets. Using single-cell RNA-seq data, cellDancer provides not only the direction of cell fate transition but also the inference of cell fate regulation including transcriptional efficiency regulation and splicing and degradation efficiency regulation which are post-transcriptional regulation. We validated our prediction of transcription, splicing,

and degradation rates by comparing them to experimental measurements (i.e., metabolic labeling sequencing).

Speaker: Dr. Kaifu Chen
Title: MEBOCOST: Metabolite-mediated Cell Communication Modeling by Single Cell Transcriptome
Abstract: We developed MEBOCOST, an algorithm for quantitatively inferring metabolite-mediated intercellular communications using single-cell RNA-seq data. The algorithm identifies cell-cell communications in which metabolites, such as lipids, are secreted by sender cells and traveled to interact with sensor proteins of receiver cells. The sensors on the receiver cell include the cell surface receptors, transporters across the cell membrane, or nuclear receptors. MEBOCOST relies on a comprehensive database of metabolite-sensor partners, which we manually curated from the literature and other public sources. MEBOCOST defines sender and receiver cells for an extracellular metabolite based on the expression levels of the enzymes and sensors, respectively, thus identifies metabolite-sensor communications between the cells. Applying MEBOCOST to mouse brown adipose tissue (BAT) successfully recaptured known metabolite-mediated cell communications and further identified new communications. Additionally, MEBOCOST identified a set of intercellular metabolite-sensor communications regulated by cold exposure in mouse BAT. MEBOCOST will be useful to researchers for investigation of metabolite-mediated cell-cell communications in many biological and disease models. The MEBOCOST software is freely available at https://github.com/zhengrongbin/MEBOCOST.

Speaker: Dr. Liang Chen (University of Southern California)
Title: Enhancing Cell-Type Identification in Single-Cell RNA-seq Data with Interpretable Deep Learning
Abstract: Identifying cell types is crucial for understanding the functional units of an organism. Machine learning has shown promise for identifying cell types, but many existing methods lack biological significance due to poor interpretability. However, it is of the utmost importance to understand what makes cells share the same function and form a specific cell type. Here, we propose CellTICS, a biologically explainable neural network for Cell-Type IdentifiCation and interpretation based on Single-cell RNA-seq data. Our CellTICS addresses this challenge by prioritizing marker genes with cell-type-specific expression, using a hierarchy of biological pathways for neural network construction, and applying a multi-predictive-layer strategy to predict cell and sub-cell types. CellTICS usually outperforms existing methods in prediction accuracy. Moreover, CellTICS can reveal pathways that define a cell type or a cell type under specific physiological conditions, such as disease or aging. The nonlinear nature of neural networks enables us to identify many novel pathways. Interestingly, some of the pathways identified by CellTICS exhibit differential expression "variability" rather than differential expression across cell types, indicating that expression stochasticity within a pathway could be an important feature characteristic of a cell type. Overall, CellTICS provides a biologically interpretable method for identifying and characterizing cell types, shedding light on the underlying pathways that define cellular heterogeneity and its role in organismal function.

Speaker: Dr. Xueqiu Lin (Stanford University)
Title: The Whole is More Than the Parts: Decoding Synergistic Networks of Multiple Non-coding Variants Linked to Cancer Risk
Abstract: 98% of the human genome is made up of non-coding regions that contain various non-coding elements. Collectively, these elements determine cell fates and disease stages. Recent GWAS studies have revealed that more than 90% of variants associated with complex diseases, such as diverse cancers, occur

in non-coding regions. In addition, the majority of the non-codding variants only have small effects. These two factors make the understanding of the function of non-coding variants one of the biggest challenges in human genomics. We all agree that" Whole is more than the sum of its parts." But what is meant by more? We used computational modeling and multiplexed CRISPRi screening to decode the synergy between synergistic regulatory elements (SREs) and explain the whole function of the enhancer network in controlling oncogene expression. Furthermore, we developed a network-based model to interpret the whole function of multiple non-coding variants in cancer risk. This work advances our understanding of enhancer biology and promotes our ability to translate the genomic information to clinical settings.

Special Session – Special Topics on Genomics and Translational Bioinformatics
Wednesday, July 19, 2023
9:30 AM – 11:10 AM
St. Petersburg I

Chairs: Ece Uzun, Shulan Tian
AMIA GenTBI Working Group

Speaker: Dr. Huihuang Yan (Mayo Clinic)
Title: Chromatin-mediated transcriptional dysregulation in T-cell Prolymphocytic Leukemia.
Abstract: T cell prolymphocytic leukemia (T-PLL) is a rare disease representing ~2% of mature lymphocytic leukemia cases in adults. T-PLL has an aggressive clinical course, with poor response to conventional chemotherapy and immunotherapy. Whole-genome and whole-exome sequencing has identified numerous structural variants, including inv(14)(q11q32) and t(14;14)(q11;q32), as well as recurrent mutations in genes involved in DNA damage response and chromatin regulation. On the other hand, RNA-seq revealed dysregulations of numerous oncogenes, including TCL1A, a core lesion in T-PLL pathogenesis. Nevertheless, the possible roles of regulatory mechanisms in T-PLL remain largely unexplored. We performed chromatin immunoprecipitation and sequencing (ChIP-seq) for histone H3 lysine 27 acetylation (H3K27ac, enhancer), H3K4 monomethylation (H3K4me1, enhancer) and trimethylation (H3K4me3, promoter), as well as RNA-seq in patients (n=6) and age-matched healthy individuals (n=3). Samples were collected with written consent and approval from the institutional review board at Mayo Clinic. Unsupervised clustering of ChIP-seq for the 3 marks and gene expression levels revealed two distinct groups corresponding to T-PLL and normal, indicating both chromatin and transcriptional reprogramming in T-PLL. There are 2,445 and 5,587 H3K27ac peaks that showed significant increase and decrease signals in T-PLL, respectively. These differential peaks are preferentially associated with differentially expressed genes that are mostly enriched in the immune response pathway. We further identified over 100 T-PLL unique super-enhancers that are associated with genes enriched in MAPK signaling, IL-2/STAT5 signaling and T-cell receptor and co-stimulatory signaling pathways. Finally, we observed up-regulation of oncogenes, such as TCL1, MYC and EZH2, that is coupled with increased occupancy of H3K27ac. Reversely, down-regulation of genes responsible for T-cell activation, including CTLA-4 and numerous signaling lymphocyte activation molecule (SLAM) family genes, and for DNA

damage response, is linked to a loss of H3K27ac and H3K4me3. Together, our analyses provide insights into the roles of epigenetic alterations in T-PLL.

Speaker: Dr. Alper Uzun (Brown University)
Title: Decoding Genomic Variations with Variant Graph Craft: A User-Friendly Tool for VCF Analysis
Abstract: The Variant Call Format (VCF) file is a highly structured text file, packed with essential genomic information such as variant positions, alleles, genotype calls, and quality scores. It has rapidly grown in popularity among researchers and clinicians for its comprehensive data richness, becoming a go-to resource for understanding genomic variations. Yet, dissecting and visually representing this data isn't straightforward—it necessitates a myriad of resources and a robust set of features. To overcome these hurdles, we developed Variant Graph Craft (VGC). This powerful tool for VCF file visualization and analysis streamlines the exploration of genetic variations. It empowers users to extract variant data, present variants visually, and create graphical representations of samples, complete with genotype information. VGC is far from just a standalone solution—it actively integrates with external resources, identfying gene function and variant frequency details across various populations. It harnesses gene function and pathway information from the Msig Database for GO terms and other sources like KEGG, Biocarta, Pathway Interaction Database, and Reactome. For detailed variant information, VGC dynamically links to gnomAD, and for pathogenic variant data, it includes ClinVar. Designed with privacy and security in mind, VGC operates on the user's local machine, bypassing the need for cloud uploads of VCF files. It supports the Human Genome Assembly Hg37, guaranteeing compatibility with a wide variety of datasets. Moreover, it offers a suite of options for genetic variation exploration, customizing the experience according to the user's specific needs by using optional phenotype input data. In essence, VGC offers a secure, user-friendly platform for investigating genetic variations. Its user-friendly features allow researchers and clinicians to effectively decipher and comprehend genomic variation data in a holistic and accessible way. Whether pinpointing specific genetic mutations or analyzing genome-wide variation patterns, VGC is an invaluable tool for anyone working in the genomics arena.VGC is freely available at https://github.com/alperuzun/VGC

Speaker: Dr. Nephi Walton (Intermountain Healthcare)
Title: Genomics and Artificial Intelligence in Clinical Care
Abstract: Our ability to build powerful artificial intelligence (AI) models on massive datasets is increasing rapidly at the same time the cost of genome sequencing is plummeting. This combination of events presents great opportunity for deploying AI with genomics in the clinical environment. In the past the massive dimensionality of genomic data combined with the lack of knowledge regarding gene function have been major barriers to the use of Artificial Intelligence (AI) in genomics. These barriers are starting to erode. With the increase in the number of population sequencing initiatives rapidly increasing and large genomic datasets becoming more widely available, the prospect of using artificial intelligence to gain insights from this data becomes more tangible and the use of such technology becomes even more important. In this session we review the current state and future directions of using AI with genomic data in a clinical environment.

Speaker: Dr. Shulan Tian (Mayo Clinic)
Title: Unified somatic calling and machine learning-based classification enhance the discovery of clonal hematopoiesis of indeterminate potential

Abstract: Clonal hematopoiesis (CH) of indeterminate potential (CHIP), driven by somatic mutations in leukemia-associated genes, confers increased risk of hematologic malignancies, cardiovascular disease, and all-cause mortality. In blood of healthy individuals, small CH clones with competitive advantage can expand over time to reach 2% variant allele frequency (VAF), the current threshold for CHIP. Nevertheless, reliable detection of low-frequency CHIP mutations requires deep targeted sequencing which is costly and not scalable. Here, we present a streamlined variant detection and refinement workflow, **UNI**fied **SO**matic calling and **M**achine learning-based classification, or UNISOM for short, to enhance CHIP discovery from whole-genome and whole-exome sequencing data that are underpowered, especially for low VAFs, due to insufficient coverage and inherent sequencing error. UNISOM utilizes a meta-caller for variant detection, which is an ensemble of three sensitive tools benchmarked over simulated genomes with CHIP spike-in, in couple with machine learning models built with variant-associated features. The model classifies annotated variants into CHIP, germline and artifact, thus minimizing the time-consuming manual inspection needed for refinement. UNISOM achieved good recall in whole-exome data, recovering nearly 80% of the CHIP mutations identified in the same cohort via deep targeted sequencing. Applied to whole-genome data from 979 individuals in the Mayo Clinic Biobank, it recapitulated the patterns previously established in much larger cohorts, including the most mutated genes, predominant mutation types and signatures, as well as strong associations of CHIP with age and smoking status. Most notably, 30% of the identified CHIP mutations had VAFs below 5%, demonstrating its high sensitivity toward small mutant clones. This workflow is applicable to the CHIP screening in population studies.

Speaker: Dr. Ece Uzun (Brown University)
Title: Unveiling the Hidden Web of Protein-Protein Interactions in Cancer and Subtypes
Abstract: Patients with the same cancer type often share a common set of protein-protein interactions (PPIs) that contribute to the disease's progression. However, natural variations within the population give rise to distinct cancer subtypes, each characterized by unique molecular mechanisms. These subtypes play a crucial role in determining differences in survival rates, response to treatment, and histological features among patients. While current treatments focus on targeting the overall cancer type, developing therapeutic strategies specific to each subtype could potentially yield improved clinical outcomes. Our understanding of the PPIs in many cancers and their subtypes remains limited. To address this knowledge gap, a novel PPI detection tool, Proteinarium has been developed by Armanious et al (Genomics 2020). Our study aims to utilize this tool to elucidate the PPIs associated with bladder, breast, colon, brain, lung, ovarian, pancreas, prostate, and thyroid cancers, along with their respective subtypes. Using the data obtained, we have developed the Atlas of Protein-Protein Interactions in Cancer (APPIC), a web-based tool for PPI visualization. Not only does APPIC enable users to visualize PPIs in specific cancer types and subtypes, but it also aggregates relevant biological and clinical information from online databases such as Human Protein Atlas, cBioportal, gProfiler and Clue.io. In our methodology, genomic data was sourced from cBioportal and analyzed using Proteinarium to categorize cancer subtypes based on PPIs. The resulting PPI data was stored in text files and formed the foundation for APPIC. The front-end of APPIC was developed using HTML, CSS, and JavaScript with the React framework, while the back-end was built using Java and JavaScript. To generate network diagrams of the PPIs, JavaScript libraries were used. API calls were made to external databases using JavaScript and Java. APPIC serves as a web application primarily focused on visualizing PPIs. Users can select a tissue of interest (e.g., "Thyroid") and a subtype (e.g., "Follicular"). APPIC generates a network diagram illustrating the PPIs and provides aggregated results from external databases. Human Protein Atlas and cBioPortal offer valuable biological and clinical information about the entire cancer subtype network, while Clue.io aids in identifying existing drug targets for specific genes,

facilitating drug repurposing efforts. Users can annotate the network using GO Terms and the KEGG database by using g:profiler. By identifying and visualizing PPIs associated with various cancer subtypes, APPIC enhances our understanding of cancer mechanisms which would potentially lead to more precise therapeutic strategies.

Concurrent Session – Genomics, Transcriptomics, Proteomics and Epigenomics I
Monday, July 17, 2023
2:00 PM – 4:35 PM
St. Petersburg II, III

Chairs: Qin Ma, Xiaojing Wang

Mitigating Heterogeneity Effects in Microbiome-based Quantitative Phenotype Prediction: A Comprehensive Workflow for Integrating Multiple Studies with Batch Normalization

Yilin Gao[1], Fengzhu Sun[1,*]
[1] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA
[*] Corresponding author: fsun@usc.edu

Abstract
Machine learning models trained on microbiome compositional data and applied on the predictions of quantitative phenotypes are increasingly used in investigating the impact of the microbiome on human health. However, such models are sensitive to heterogeneity effects, which can negatively impact reproducibility and generalizability across different studies. Here, we investigate the effectiveness of different normalization and integration methods in mitigating heterogeneity effects on cross-study predictions of quantitative phenotypes. Using simulations and real data applications, we show that normalization, particularly ComBat and ConQuR methods, combined with integration methods can effectively remove heterogeneity effects and improve prediction performance. Interestingly, we find that the commonly used merging method may not be optimal in the context of quantitative phenotype prediction. Overall, our study highlights the importance of mitigating heterogeneity effects on machine learning model reproducibility in the area of quantitative phenotype, leading the direction towards more effective cross-study predictions.

Comprehensive Cross Cancer Analyses Reveal Mutational Signature Cancer Specificity

Rui Xin[1], Limin Jiang[2], Hui Yu[2], Jijun Tang[1*], Yan Guo[2*]
[1] Department of Computer Science, University of South Carolina, Columbia, SC 29201, USA
[2] Department of Public Health and Sciences, Sylvester Comprehensive Cancer Center, University of Miami, Miami, 33136, USA
[*] Corresponding authors

Abstract
Mutational signatures refer to distinct patterns of DNA mutations that occur in a specific context or under certain conditions. It is a powerful tool to describe cancer etiology. We conducted a study to show cancer

heterogeneity and cancer specificity from the aspect of mutational signatures through collinearity analysis and machine learning techniques. The results show that while the majority of the mutational signatures are distinct, similarities between certain mutational signature pairs can be observed through both mutation patterns and mutational signature abundance. The observation can potentially assist to determine the etiology of yet elusive mutational signatures. Further analysis using machine learning approaches demonstrated moderate mutational signature cancer specificity. Skin cancer among all cancer types demonstrated the strongest mutational signature specificity.

A Weighted Two-stage Sequence Alignment Framework to Identify DNA Motifs from ChIP-exo Data

Yang Li[1,†], Yizhong Wang[2,†], Cankun Wang[1], Anne Fennell[3], Anjun Ma[1], Jing Jiang[1], Zhaoqian Liu[1,2], Qin Ma[1,4,*], and Bingqiang Liu[2,*]

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA, 43210,

[2] School of Mathematics, Shandong University, Jinan, China, 250100,

[3] Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, SD, 57007, USA,

[4] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, 43210, USA,

[†] These authors contributed equally: Yang Li, Yizhong Wang.

[*] To whom correspondence should be addressed. bingqiang@sdu.edu.cn Qin.Ma@osumc.edu

ABSTRACT

Identifying precise transcription factor binding sites (TFBS) or regulatory DNA motifs plays a fundamental role in researching transcriptional regulatory mechanisms in cells and in helping construct regulatory networks. Current algorithms developed for motif searching focus on the analysis of ChIP-enriched peaks but are not able to integrate the ChIP signal in nucleotide resolution. We present a weighted two-stage alignment tool (TESA). Our framework implements an analysis workflow from experimental datasets to TFBS prediction results. It employs a binomial distribution model and graph searching model with ChIP-exonuclease (ChIP-exo) reads depth and sequence data. TESA can effectively measure the possibility for each position to be an actual TFBS in a given promoter sequence and predict statistically significant TFBS sequence segments. The algorithm substantially improves prediction accuracy and extends the scope of applicability of existing approaches. We apply the framework to a collection of 20 ChIP-exo datasets of E. coli from proChIPdb and evaluate the prediction performance through comparison with three existing programs. The performance evaluation against the compared programs indicates that TESA is more accurate for identifying regulatory motifs in prokaryotic genomes.

A mouse-specific model to detect genes under selection in tumors

Hai Chen[1,2], Jingmin Shu[1,2], Li Liu[1,2,*]

[1] College of Health Solutions, Arizona State University, Phoenix, AZ, 85004, USA

[2] Biodesign Institute, Arizona State University, Tempe, AZ, 85281, USA

[*] To whom correspondence should be addressed (liliu@asu.edu)

Abstract

Mouse is a widely used model organism in cancer research. However, no computational methods exist to identify cancer driver genes in mice due to a lack of labeled training data. To address this knowledge gap, we adapted the GUST (genes under selection in tumors) model, originally trained on human exomes, to mouse exomes using transfer learning. The resulting tool, called GUST-mouse, can estimate long-term and short-term evolutionary selection in mouse tumors, and distinguish between oncogenes, tumor suppressor genes, and passenger genes using high throughput sequencing data. We applied GUST-mouse to analyze 65 exomes of mouse primary breast cancer models, leading to the discovery of 24 driver genes. The GUST-mouse method is available as an open-source R package on github (https://github.com/liliulab/gust.mouse).

A machine learning pipeline to detect open chromatin regions from cfDNA sequencing data

Yuxin Liu[1], Yuqian Liu[1], Xiaoyan Zhu[1], Jiayi Ren[1], Xin Lai[1], Xuanping Zhang[1], Jiayin Wang[1]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China

Abstract. Open chromatin regions (OCRs) are specialized regions in the human genome that can be accessed by DNA regulatory elements. Detection of OCRs could help to understand the mechanisms of human disease. Traditional chro-matin open area detection methods mainly use tissue cells or cell lines cultured in vitro, which makes it difficult to obtain cell samples, complicated experi-mental steps, and difficult to be applied in clinical scenarios. Recent studies have shown that the characteristic signal of cfDNA is associated with chromatin accessibility, and cfDNA in plasma is easy to obtain and less harmful. At pre-sent, OCR detection based on cfDNA-seq data generally needs to obtain possi-ble open region locations in advance, or use a fixed length detection interval for detection, which is unable to obtain more accurate open region locations. There-fore, it is of significance to explore a method for OCR detection based on nu-cleosome interval. This paper proposes a new method for detecting open chro-matin regions based on cfDNA sequencing data. The method uses human plas-ma cfDNA sequencing data, calculates WPS waveform and cfDNA fragment length distribution matrix, converts the original data into image data, adjusts the image size based on data characteristics, and designs a new attentional mecha-nism module to help the model obtain important information from simple imag-es. The improved noise processing method was used to remove the noise data which was difficult to distinguish by traditional methods, and finally the OCR set with nucleosome interval as unit was obtained. We compared the percentage overlap between our OCR and the OCR obtained by other methods. The results showed that more than 90% of the TSS regions in housekeeping genes were de-tected, and at relatively high enrichment degree, the overlap rate between our OCRs and ATAC-seq or DNase-seq was greater than 70%, indicating the effec-tiveness of our method.

Keywords: Sequencing data analysis; cell-free DNA, open chromatin region; detection algorithm; machine learning approach.

Detection of viral infection in cell lines using ViralCellDetector

Rama Shankar[1,*], Shreya Paithankar[1], Suchir Gupta[1], Bin Chen[1,2,3*]
[1] Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI 49503, USA
[2] Department of Pharmacology and Toxicology, College of Human Medicine, Michigan State University, Grand Rapids, Michigan, USA

[3] Department of Computer Science and Engineering, College of Engineering, Michigan State University, East Lansing, Michigan, USA
[*] To whom correspondence should be addressed: RS: ramashan@msu.edu and BC: chenbi12@msu.edu.

ABSTRACT
Cell lines are commonly used in research to study biology, including gene expression regulation, cancer progression, and drug responses. However, cross-contaminations with bacteria, mycoplasma, and viruses are common issues in cell line experiments. Detection of bacteria and mycoplasma infections in cell lines is relatively easy but identifying viral infections in cell lines is difficult. Currently, there are no established methods or tools available for detecting viral infections in cell lines. To address this challenge, we developed a tool called ViralCellDetector that detects viruses through mapping RNA-seq data to a library of virus genome. Using this tool, we observed that around 10% of experiments with the MCF7 cell line were likely infected with viruses. Furthermore, to facilitate the detection of samples with unknown sources of viral infection, we identified the differentially expressed genes involved in viral infection from two different cell lines and used these genes in a machine learning approach to classify infected samples based on the host response gene expression biomarkers. Our model reclassifies the infected and non-infected samples with an AUC of 0.91 and an accuracy of 0.93. Overall, our mapping- and marker-based approaches can detect viral infections in any cell line simply based on readily accessible RNA-seq data, allowing researchers to avoid the use of unintentionally infected cell lines in their studies.

Key words: Cell lines, Viral infection, Bacterial infection, Differentially expressed genes, RNA-seq data, Random Forest, and Machine learning.

A comprehensive benchmark of transcriptomic biomarkers for immune checkpoint blockades

Hongen Kang[1,2], Xiuli Zhu[1,2], Ying Cui[1,2], Zhuang Xiong[2,3], Wenting Zong[2,3], Yiming Bao[2,3,*], Peilin Jia[1,2,3,*]
[1] CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
[*] To whom correspondence should be addressed: Dr. Peilin Jia, pjia@big.ac.cn; Dr. Yiming Bao, baoym@big.ac.cn

ABSTRACT
Immune checkpoint blockades (ICB) have revolutionized cancer therapy by inducing durable clinical responses, but only a small percentage of patients can benefit from ICB treatments. Many studies have established various biomarkers to predict ICB responses. However, different biomarkers were found with diverse performances in practice, and a timely and unbiased assessment has yet to be conducted due to the complexity of ICB-related studies and trials. In this study, we manually curated 29 published datasets with matched transcriptome and clinical data from more than 1400 patients, and uniformly preprocessed these datasets for further analyses. In addition, we collected 39 transcriptomic biomarkers, and based on the nature of the corresponding computational methods, we categorized them into the gene-set-like group (with the self-contained design and the competitive design, respectively) and the deconvolution-like group. Next, we

investigated the correlations and patterns of these biomarkers and utilized a standardized workflow to systematically evaluate their performance in predicting ICB responses and survival statuses across different datasets, cancer types, antibodies, biopsy times, and combinatory treatments. In our benchmark, most biomarkers showed poor performance in terms of stability and robustness across different datasets. Two scores (TIDE and CYT) had a competitive performance for ICB response prediction, and two others (PASS-ON and EIGS_ssGSEA) showed the best association with clinical outcome. Finally, we developed ICB-Portal (https://ngdc.cncb.ac.cn/icb) to host the datasets, biomarkers, and benchmark results and to implement the computational methods for researchers to test their custom biomarkers. Our work provided valuable resources and a one-stop solution to facilitate ICB-related research.

Keywords: Immune checkpoint blockade (ICB), Transcriptomic biomarkers, Benchmark, ICB-Portal, TIDE

Concurrent Session – Genomics, Transcriptomics, Proteomics and Epigenomics II
Tuesday, July 18, 2023
9:30 AM – 12:05 PM
Williams/Demens

Chairs: Renzhi Cao, Jing Wang

AlphaCluster: Coevolutionary driven residue-residue interaction models enable quantifiable clustering analysis of de novo variants to enhance predictions of pathogenicity

Joseph Obiajulu[1,2], Ranger Kuang[3], Lesi He[4], Guoije Zhong[2], Jake Hagen[1,2], Chang Shu[1,2], Wendy K. Chung[1, #], Yufeng Shen[2,4,#]
[1] Department of Pediatrics, Columbia University, New York, NY, USA
[2] Department of Systems Biology, Columbia University, New York, NY USA
[3] The Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY, USA
[4] Department of Biostatistics, Columbia University, New York, NY USA
[5] Department of Biomedical Informatics, Columbia University, New York, NY, USA
[#] Corresponding authors: W.K.C (wkc15@columbia.edu) and Y.S. (ys2411@cumc.columbia.edu)

Abstract
Missense variants have highly variable effects and effect size, which often makes it challenging to distinguish pathogenic and non-pathogenic variants and subsequently implicate new genes for disease association in studies of de novo and inherited rare variants. Importantly, missense variants can be the sole molecular mechanism for some genetic disorders, and so statistical approaches tailored for the analysis of missense variants are critical. Analysis of the clustering of missense variants is a promising approach which leverages the fact that missense variants in protein domains often have similar effects on function. Here we describe a new clustering analysis approach, AlphaCluster, a statistical method which quantifiably analyzes the spatial clustering of de novo variants by mapping missense residues onto the protein tertiary structure. We show that our approach can quantify the evidence supporting

pathogenic missense variants and increase the power to detect clustering when compared to available genomic clustering tools. Using AlphaCluster, we identified genes newly implicated in autism spectrum disorder and neurodevelopmental disorders (NDD). We also apply AlphaCluster to protein complexes and detect an association between the gamma aminobutyric acid receptor complex (GABA-A $\alpha 1\beta 2\gamma 2$ receptor).

Mutation Density Analyses on Long Noncoding RNA Reveal Comparable Patterns to Protein-Coding RNA and Prognostic Value

Chaoyi Troy Zhang[1*], Hui Yu[1*], Yongsheng Bai[2], Yan Guo[1]
[1] Department of Public Health and Sciences, Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL 33136
[2] Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197
[*] Equal contribution

Abstract
Mutations and gene expression are the two most studied genomic features in cancer research. In the last decade, the combined advances in genomic technology and computational algorithms broadened mutation research with the concept of mutation density and expanded the traditional scope of protein-coding RNA to noncoding RNAs. However, mutation density analysis was never integrated with non-coding RNAs. In this study, we examined long non-coding RNA (lncRNA) mutation density patterns of 57 unique cancer types using 80 cancer cohorts. Our analysis revealed that lncRNAs exhibit similar mutation density patterns as protein-coding RNAs. These patterns include mutation peak and dip around translation start sites of lncRNA. In many cohorts, these patterns demonstrate statistically significant transcription strand bias. We further quantified transcription strand biases and showed that some of these biases are associated with patient prognosis. The prognostic effect may be exerted due to strong DNA repair mechanisms associated with the individual patient. For the first time, our study combined mutational density patterns with lncRNA mutations, and the results demonstrated remarkably comparable patterns between protein-coding RNA and lncRNA, further illustrating lncRNA's potential roles in cancer research.

Systematic assessment of small RNA profiling in human extracellular vesicles

Jing Wang[†1,2], Hua-chang Chen[†1,2], Quanhu Sheng[1,2], Renee Dawson[3], Robert J. Coffey[3,4], James G. Patton[5], Alissa M. Weaver[3,6], Yu Shyr[*1,2], Qi Liu[*1,2]
[1] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37232, USA [2] Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[3] Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[4] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[5] Department of Biological Sciences, Vanderbilt University, Nashville, TN 37232, USA
[6] Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[†] These authors contributed equally.

* Corresponding author: yu.shyr@vumc.org; qi.liu@vumc.org

Abstract

Motivation: Extracellular vesicles (EVs) are produced and released by most cells and are now recognized to play a role in intercellular communication through the delivery of molecular cargo, including proteins, lipids, and RNA. Small RNA sequencing (small RNA-seq) has been widely used to characterize the small RNA content in EVs. However, there is lack of a systematic assessment of the quality, technical biases, RNA composition, and RNA biotypes enrichment for small RNA profiling of EVs across cell types, biofluids, and conditions.

Results: We collected and reanalyzed small RNA-seq datasets for 2,799 samples from 83 studies involving 55 with EVs only and 28 with both EVs and matched donor cells. We assessed their quality by the total number of reads after adapter trimming, the overall alignment rate to the host and non-host genomes, and the proportional abundance of total small RNA and specific biotypes, such as miRNA, tRNA, rRNA, and Y RNA. We found EV extraction methods varied in reproducibility with effects on small RNA composition. Comparing proportional abundances of RNA biotypes between EVs and matched donor cells, we discovered rRNA and tRNA fragments relatively enriched but miRNAs and snoRNA depleted in EVs. Except the export of eight miRNAs being context-independent, the selective release of most miRNAs into EVs were study-specific.

The genetic regulation of the biogenesis of human isomiRs

Guanglong Jiang[1,2], Jill L. Reiter[2], Chuanpeng Dong[1], Yue Wang[2], Fang Fang[2], Zhaoyang Jiang[3], Yunlong Liu[1,2]

[1] Department of BioHealth Informatics, Indiana University Purdue University Indianapolis, Indianapolis, IN, USA

[2] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA

[3] Department of Computer Science, Purdue University, West Lafayette, IN, USA

Abstract

MicroRNA plays a critical role in regulating gene expression post-transcriptionally. Variations in mature microRNA sequences, known as isomiRs, arise from imprecise cleavage, nucleotides substitution or addition. These isomiRs can compete with their canonical counterparts or target different mRNAs, thereby expanding their scope in post-transcriptional regulations. Our study investigated the relationship between cis-acting single nucleotide polymorphisms (SNPs) in precursor miRNA regions and isomiR composition, represented by the ratio of a specific 5'-isomiR subtype to all isomiRs identified for a particular mature miRNA. The significant associations between 136 SNP-isomiR pairs were identified in the study. Of note, rs6505162 was significantly associated with both 5'-extension of hsa-miR-423-3p and 5'-trimming of hsa-miR-423-5p. Comparison of breast cancer and normal samples revealed that both isomiRs were significantly higher in tumors than in normal tissue. This study sheds light on the genetic regulation of isomiR maturation and advances our understanding of post-transcriptional regulation by microRNA.

SynthQA - Hierarchical Machine Learning-based Protein Quality Assessment

Mikhail Korovnik[1*], Sheng Wang[2†], Junyong Zhu[2‡], Kyle Hippe[1§], Jie Hou[3¶], Dong Si4[**], Kiyomi Kishaba5[††], Renzhi Cao[1‡‡]

[1] Department of Computer Science, Pacific Lutheran University, Tacoma,WA, 98447,USA
[2] Department of Computer Science and Technology, Anhui University, Hefei, Anhui, 230601, China
[3] Department of Computer Science, Saint Louis University, Saint. Louis, MO, 63103, USA
[4] Division of Computing and Software Systems, University ofWashington Bothell, Bothell,WA, 98011, USA
[5] Department of Humanities, Pacific Lutheran University, Tacoma,WA, 98447, USA
[‡‡] Correspondence: Renzhi Cao (caora@plu.edu)

Abstract:
It has been a challenge for biologists to determine 3D shapes of proteins froma linear chain of amino acids. Experimental techniques, such as X-ray crystallography or Nuclear Magnetic Resonance, are time-consuming. This highlights the importance of computational methods for protein structure predictions. In the field of protein structure prediction, ranking the predicted protein decoys and selecting the one closest to the native structure is known as protein model quality assessment (QA). Traditional QA methods don't consider different types of features from the protein decoy, lack various features for training machine learning models, and don't consider the relationship between features. In this research, we introduce a new single-model QA tool SynthQA that incorporates multi-scale features from protein structures, utilizes the hierarchical architecture of training machine learning models, and predicts the quality of any protein decoy. Based on our experiment, the new hierarchical architecture improves upon the accuracy of traditional machine learning-based methods. It also considers the relationship between features and generates more features to improve accuracy of machine learning techniques.

Characterizing protein structural features of alternative splicing and isoforms using AlphaFold 2

Yuntao Yang[1], Yuhan Xie[2], Zhao Li[1], Chiamaka S. Diala[1], Meer A. Ali[1], Rongbin Li[1], Yi Xu[1], Sayed-Rzgar Hosseini[1], Erfei Bi[3], Hongyu Zhao[2], W. Jim Zheng[1]
[1] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.
[2] Department of Biostatistics, Yale University School of Public Health, New Haven, CT, USA.
[3] Department of Cell and Developmental Biology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

Abstract
Alternative splicing is a cellular process in eukaryotes, which modifies pre-mRNA and generates multiple protein isoforms for a single gene. Our current knowledge of alternative splicing events in the context of protein structures is very limited due mainly to a lack of sufficient protein structural data. In this study, we employed AlphaFold 2, the recent breakthrough in protein structure prediction, to perform a thorough analysis of the entire spectrum of alternative splicing for ~3,500 human genes, which has enabled us to gain

a deeper understanding of the process at the protein structural level. Our analysis systematically characterized structural features in alternatively spliced regions and identified structural changes after alternative splicing events. Our study revealed that alternative splicing tends to change the structure of residues located mainly in coils and beta-sheets. Moreover, we found that intrinsically disordered and highly exposed residues are significantly enriched in human alternatively spliced regions. Specifically, our study on the SEPTIN9 protein has revealed the possible involvement of intrinsically disordered regions in alternative splicing and its evolution. Furthermore, we uncovered two missense mutations in the Tau protein that could alter alternative splicing and potentially contribute to the pathogenesis of Alzheimer's disease. Thus, by comprehensive statistical analysis of extensive protein structural data, our work sheds new light on the relationship between alternative splicing, evolution, and human disease.

Keywords: alternative splicing, protein isoforms, protein structure, AlphaFold 2

Concurrent Session – Medical Informatics, Public Health Informatics and Pharmacoinformatics I
Monday, July 17, 2023
1:40 PM – 4:35 PM
Williams/Demens

Chairs: Mei Liu, Satish Mahadevan Srinivasan

The association between nonalcoholic fatty liver disease (NAFLD) status and physical exam or biochemical parameters

Weiru Han,[1,$] Tianrui Zhu,[2,$] Zhengli Tang,[3,$] Robert Morris,[2,$] Kun Bu,[1] Fang Wang,[3] Lin Fan,[3] Weijian Wang,[3,*] Yiming Hao,[4,*] Yiqin Wang,[4,*] and Feng Cheng[2,5,*]

[1]Department of Mathematics & Statistics, College of Art and Science, University of South Florida, Tampa, FL 33620, USA

[2]Department of Pharmaceutical Sciences, Taneja College of Pharmacy, University of South Florida, Tampa, FL 33612, USA

[3]Shuguang Hospital Health Examination Center Affiliated with Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

[4]Shanghai Key Laboratory of Health Identification and Assessment/Laboratory of Traditional Chinese Medicine Four Diagnostic Information, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

[5]Department of Biostatistics and Epidemiology, College of Public Health, University of South Florida, Tampa, FL 33612, USA

[$] These authors contributed equally to this work.

Correspondence should be addressed to Weijian Wang (wwjhxj@163.com), Yiming Hao (hymjj888@163.com) Yiqin Wang (wangyiqin2380@sina.com) and Feng Cheng (fcheng1@usf.edu)

Abstract

Introduction: Nonalcoholic Fatty liver disease (NAFLD) is a condition in which there is an excessive accumulation of fat in liver cells. The consequences of high levels of intrahepatic fat can be serious, and it's crucial to identify possible factors that are strongly associated with the disease to provide a better medical intervention in the future.
Objectives: The objective is to identify the optimal algorithms for predicting NAFLD. Ranking the importance of features and identifying potential factors that have a strong relationship with NAFLD are also of interest.
Methods: The data was provided by the Shuguang Hospital affiliated with Shanghai University of Traditional Chinese Medicine. After preprocessing, the data contained 5479 records with 18 physical exam or biochemical parameters. Six algorithms were compared, which are Linear discriminant analysis (LDA), generalized linear model (GLM), K-nearest neighbor (KNN), naive Bayes, stochastic gradient boosting (SGB), and random forest (RF). The evaluation metrics included Youden's index and the Area under the ROC Curve (AUC) value. Furthermore, we divided the data into 5 age groups and applied the chosen model to evaluate the performance and feature importance in different age groups.
Results: SGB was found to be the most accurate machine learning model for prediction of NAFLD status. The algorithm exhibited the highest performance for the youngest age group (under 36 years old). The performance gradually decreased with age.Using the optimal SGB model, BMI was identified as the strongest predictor of NAFLD status across all 5 age groups. Additional parameters comprising the top strongest predictors were body mass index (BMI), triglceride (TG), and serum gamma-glutamyl transferase (γ-GT). HDL cholesterol, serum creatinine (Cr), and fasting plasma glucose (FPG) exhibited notable temporal patterns across each age group. The strength of FPG as a predictor increased with respect to age while HDL cholesterol prediction strength generally decreased with age.
Conclusion: SGB was found to be the most accurate machine learning model for prediction of NAFLD status. The association between NAFLD status and some biochemical parameters may be heavily mediated by patient age and should be considered when predicting NAFLD risk.

Keywords: Machine learning, stochastic gradient boosting, BMI, triglceride, creatinine, HDL, FPG, temporal pattern, gamma-glutamyl transferase

---

Behavioral and demographic profiles of HIV contact networks in Florida

Yiyang Liu[*1], Christina Parisi[*1], Rebecca Fisk-Hoffman[1], Marco Salemi[2], Diego Viteri1, Mattia Prosperi[1], and Simone Marini[1,2,§]

[1] Department of Epidemiology, University of Florida
[2] Department of Pathology, University of Florida
[*] equal contribution
[§] corresponding author: simone.marini@ufl.edu

Abstract
Introduction: To complete the Ending the HIV Epidemic initiative in areas with high HIV incidence, there needs to be a greater understanding of the demographic, behavioral, and geographic factors that influence the rate of new HIV diagnoses. This information will aid the creation of targeted prevention and intervention efforts. The aim of this study is to identify the geographic distribution of risk groups and their role within potential transmission networks in Florida.

Methods: Public data from the Florida Department of Health and behavioral data from the Surveillance Tools and Reporting System (STARS) between 2012-2022 were used in these analyses. We analyzed records as a combination of variables of interest (gender, age, race/ethnicity, and HIV risk group) in order to create demographic-behavioral profiles (DBPs) that represent the profiles of newly-diagnosed people with HIV. We then used the resulting DBPs to characterize Florida counties and HIV Coordination Areas and calculated the county-to-county and area-to-area rank (Spearman) correlation. We then drew a dendrogram based on the correlation matrix and identified clusters of similar counties and areas.
Results: We identified 37 DBPs. The largest DBPs were Hispanic and non-Hispanic Black males aged 25-49 reporting male-to-male sexual contact (MMSC), non-Hispanic White males aged 25-49 reporting MMSC, and non-Hispanic Black females aged 25-49 reporting heterosexual contact. The state could be broken up generally into two transmission clusters by region: Northwestern/Northern and Central/Southern. We identified several counties with similar DBPs that were not in the same HIV Coordination Area.
Conclusion: We identified distinct risk groups and clusters of transmission throughout Florida. These results can help regions identify health disparities and allocate their HIV prevention and intervention resources accordingly. The goal of this work was to highlight areas of need in a high incidence setting, not contribute to existing stigma against vulnerable groups, and it is important to consider the ethics and possible harm of advanced methodologies such as contact network analysis when addressing public health problems.

Association between ABCG1/TCF7L2 and type 2 diabetes mellitus：An intervention trial based case-control study

Yinxia Su[1#] Xiangtao Liu[1#] Conghui Hui[2] Hua Yao[3*]
[1] School of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, Xinjiang, China;
[2] School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang, China;
[3] School of Health Management, Xinjiang Medical University, Urumqi, Xinjiang, China.
Presenter's email address: yinxia_su@xjmu.edu.cn

Abstract
Background and Objective： Type 2 diabetes mellitus (T2DM) is the result of both genetic and environmental factors. Environmental factors may contribute to the occurrence and development of T2DM by influencing epigenetic modification. DNA methylation is a major modification mode and an important regulatory mechanism of epigenetic inheritance, which is considered to be an important phenotypic outcome and marker of disease progression. In this study, we focused on the methylation sites of TCF7L2 and ABCG1 genes that are most strongly associated with T2DM. By conducting intervention experiments in Uyghur population, which has been less studied, to analyze the potential functions of SNP-CG sites rs7901695 and cg06500161 of the above two genes as biomarkers in the development of T2DM, and provide evidence for personalized health management of T2DM in Uyghur people.
Methods: 320 patients with T2DM and 332 patients without T2DM were treated with dietary pagoda-based health education intervention. The demographic data and basic physical biochemical indexes before and after intervention were collected by questionnaire survey and physical biochemical examination. SNP typing was performed by Taqman-MGB probe method, and gene methylation was detected by pyrosequencing method.
Results:
1. The genotypes of SNP sites corresponding to the methylation site cg06500161 of ABCG1 gene were all

CC type without gene polymorphism, so the polymorphism of this gene locus was not analyzed. The rs7901695 genotypes of TCF7L2 gene were TT, TC and CC. But only 98 out of 332 samples contained the C allele, and the methylation modification was limited to cytosine in GC sequence. Due to the small sample size, the correlation analysis between methylation level and T2DM at this site was not conducted.
2. The rs7901695 genotype difference of TCF7L2 was statistically significant between the case group and the control group ($P < 0.05$). After adjusting for covariates (smoking, alcohol consumption, exercise, FPG, obesity and hypertension), genotype of rs7901695 in TCF7L2 gene was associated with genetic susceptibility to T2DM in addition (TC vs TT, $P = 0.047$; CC vs TT, $P = 0.010$), dominant ($P = 0.015$) and recessive ($P = 0.039$) models.
3. Before intervention, there were significant differences in the intake of water between the case group and the control group ($P < 0.05$); After intervention, there was statistical significance in the intake of coarse grains, fruits, aquatic products, eggs, dairy products, soy products, nuts, edible oils and water between the case group and the control group ($Ps < 0.05$). Logistic regression analysis showed that methylation of ABCG1 gene was correlated with T2DM susceptibility after adjustment of covariable before intervention ($P = 0.015$, OR: 1.023; 95%CI: 1.004~1.041) but not after intervention.
4. Generalized Multifactor Dimensionality Reduction (GMDR) showed rs7901695 locus of TCF7L2 gene and cg06500161 locus of ABCG1 gene had interaction with hypertension, dyslipidemia, abdominal obesity and obesity, and also had interaction with drinking, smoking and exercise.
Conclusions: The polymorphism of rs7901695 site of TCF7L2 gene is associated with the incidence of T2DM in Uyghurs. The interaction of rs7901695 site of TCF7L2 gene and cg06500161 site of ABCG1 gene with environmental factors may increase the risk of T2DM in Uyghurs. The interaction between cg06500161 site of ABCG1 gene and environmental factors on T2DM varied with the intervention. Cg06500161 site of ABCG1 may serve as a biomarker to evaluate the effect of T2DM interventions.

Keywords: ABCG1; TCF7L2; Single nucleotide polymorphism; Methylation; Type 2 diabetes mellitus

---

Smoothing spline analysis of variance models: A new tool for the analysis of accelerometer data

Rui Xie[1,*], Lulu Chen[1,*], Joon-Hyuk Park[2], Jeffrey Stout[3], Ladda Thiamwong[4]
[1] Department of Statistics and Data Science University of Central Florida
[2] School of Kinesiology and Rehabilitation Sciences, University of Central Florida
[3] College of Nursing, University of Central Florida
[4] Department of Mechanical and Aerospace Engineering, University of Central Florida

Abstract
Accelerometer data is commonplace in physical activity research, exercise science, and public health studies, where the goal is to understand and compare physical activity differences between groups and/or subject populations, and to identify patterns and trends in physical activity behavior to inform interventions for improving public health. We propose using mixed-effects smoothing spline analysis of variance (SSANOVA) as a new tool for analyzing accelerometer data. By representing data as functions or curves, smoothing spline allows for accurate modeling of the underlying physical activity patterns throughout the day, especially when the accelerometer data is continuous and sampled at high frequency. The SSANOVA framework makes it possible to decompose the estimated function into the portion that is common across groups (i.e., the average activity) and the portion that differs across groups. By decomposing the function

of physical activity measurements in such a manner, we can estimate group differences and identify the regions of difference. In this study, we demonstrate the advantages of utilizing SSANOVA models to analyze accelerometer-based physical activity data collected from community-dwelling older adults across various fall risk categories. Using Bayesian confidence intervals, the SSANOVA results can be used to reliably quantify physical activity differences between fall risk groups and identify the time regions that differ throughout the day.

Keywords: Smoothing spline ANOVA, Functional data analysis, Accelerometer data, Physical activity, Mobile health, and Wearable devices.

Exploring Drug-drug Interaction Information from PubMed using Association Rules

Kun Bu [1,$], Weiru Han [1,$], Robert Morris [2,$], and Feng Cheng [2,3,*]
[1] Department of Mathematics & Statistics, College of Art and Science, University of South Florida, Tampa, FL 33620, USA
[2] Department of Pharmaceutical Science, Taneja College of Pharmacy, University of South Florida, Tampa, FL 33613, USA
[3] Department of Biostatistics & Epidemiology, College of Public Health, University of South Florida, Tampa, FL 33613, USA
[$] These authors contributed equally to this work.
[*] Correspondence to: Feng Cheng (fcheng1@usf.edu)

Abstract
As the literature pertaining to drug-drug interactions (DDIs) increases, the complexity and discernibility of the information increases as well. In this study, a novel application of association rule mining were utilized to generate models that accurately identify DDI-related terms based on keywords and MeSH terms of PubMed literature. Four different drugs including warfarin, theophylline, rifampin, and cyclosporine were used as test sets for the accuracy of these association rule models, which had AUC values of 0.90, 0.81, 0.84, and 1.0 respectively. Additionally, co-occurrence heatmaps and directed acyclic graphs were used to show relationships between protein and drug terms related to DDIs. This systematic methodology allows for faster surveying of existing literature and allows for easier identification and display of relevant drug and protein terms pertaining to DDIs of interest.

Keywords: DDI, association rule mining (ARM), easyPubMed, warfarin, cyclosporine, rifampin, theophylline, MeSH, co-occurrence, heatmap

---

Pan-cancer mutational signature surveys correlated cancer racial disparities with geospatial environmental exposures, and viral infections

Judy Bai[1], Katherine Ma[1], Shangyang Xia[1], Richard Geng[1], Limin Jiang[1], Xi Gong[3], Hui Yu[1], Shuguang Leng[2], Yan Guo[1*]
[1] Department of Public Health and Sciences, Sylvester Comprehensive Cancer Center, University of Miami, FL, 33136, USA
[2] Comprehensive Cancer Center, Albuquerque, University of New Mexico, NM, 87109, USA
[3] Geography & Environmental Studies, University of New Mexico, Albuquerque, NM, 87109, USA

*Corresponding author: Yan Guo (yanguo1978@gmail.com)

Abstract

Background: Cancer has been disproportionally affecting minorities due to socioeconomic, environmental, and genetic disparities. Genomic-based cancer disparity analyses have been less common. In the past decade, mutational signatures were established as the characteristic footprints of endogenous or exogenous carcinogens. The disparities caused by uneven exposures to environmental pollutants or viral infections may be recorded in mutational signatures.

Methods: Utilizing mutation dataset generated from a large cancer consortium, we were able to explore mutational signatures from the aspect of racial disparity concerning geospatial environmental exposures in the form of 449 air pollutants modeled and archived by the United States Environmental Protection Agency from 2007-2017 and hepatitis B and C viruses and human papillomavirus infection status.

Results: Mutation frequencies of key oncogenic genes varied substantially between different races which are translated into disparity in mutational signatures. Eleven mutational signatures were found to be racially different. Particularly, aflatoxin, an affirmed carcinogen for liver cancer is found to be higher in Asian liver cancer patients than in White patients (adjusted P = 0.002). In-depth analyses revealed that aflatoxin mutational signature is exacerbated by hepatitis infection for Asian patients (HBV P = 0.006, HCV P = 0.004), but not for White patients, which suggests predisposed genetic or genomic disadvantage for Asians concerning aflatoxin.

Conclusions: Environmental pollution exposures negatively affect cancer patients by increasing mutational burden and worsening cancer prognosis.This study demonstrates the representative links between mutational signatures and carcinogen exposures including chemical pollutants and oncovirus infections.

Characterizing Diseases using Genetic and Clinical Variables: A Data Analytics Approach

Madhuri Gollapalli [*1], Harsh Anand [†1,2], and Satish Mahadevan Srinivasan [‡1]

[1] School of Graduate Professional Studies, Penn State Great Valley, Malvern, PA, USA

[2] Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA

[*] madhuri1gollapalli@gmail.com

[†] yyf8rj@virginia.edu

[‡] sus64@psu.edu

Abstract

Background: In the era of big data, predictive analytics plays a vital role in precision medicine aimed at personalized patient care. To aid in precision medicine, the aim of the study is to iden-tify a subset of genetic and clinical variables that can serve as predictors for classifying diseased tissue/disease types. To achieve this, experiments were carried out on a set of diseased tissues obtained from the L1000 dataset to understand the differences in the functionality and predictive capabilities of genetic and clinical variables.

Results: The results showed that landmark genes were slightly better in clustering diseased tissue types when compared to any random set of 978 non-landmark genes and the difference is statistically significant. Furthermore, it was evident that both clinical and genetic variables were important in predicting the diseased tissue types. The top three clinical predictors for predicting diseased tissue types were identified as Morphology, Gender, and the Age of Diagnosis. Additionally, this study explored the possibility of using the latent representations of the clusters of landmark and non-landmark genes as predictors for a

Multinomial Logistic Regression (MLR) classifier. The classification models built using MLR revealed that landmark genes have the capability to serve as a subset of genetic variables and/or as a proxy for clinical variables.
Conclusion: This study concludes that the use of predictive analytics and dimensionality re-duction techniques can help identify relevant predictors for use in precision medicine, ultimately resulting in more accurate predicting of the disease types.

Keywords: tissue classification, L1000 dataset analysis, landmark genes, non-landmark genes, multinomial logistic regression

---

The Association between Warfarin usage and International normalized ratio increase: Systematic analysis of FDA Adverse Event Reporting System (FAERS)

Robert Morris [1,3,$] , Megan Todd [1,$] , Nicole Zapata Aponte [1,$], Milagros Salcedo [1,$], Matthew Bruckner [1,$], Alfredo Suarez Garcia [1,$], Rachel Webb [1,$], Kun Bu [2], Weiru Han [2], Feng Cheng [1,3,*]
[1] Department of Pharmaceutical Sciences, Taneja College of Pharmacy, University of South Florida, Tampa, FL 33612, USA
[2] Department of Mathematics & Statistics, College of Art and Science, University of South Florida, Tampa, FL 33620, USA
[3] Department of Biostatistics and Epidemiology, College of Public Health, University of South Florida, Tampa, FL 33612, USA
[$] These authors contributed equally to this work.
Correspondence should be addressed to Feng Cheng; fcheng1@usf.edu

Abstract
Background: Elevated international normalized ratio (INR) has been commonly reported as an adverse drug event (ADE) for patients taking warfarin for anticoagulant therapy.
Objective: The purpose of this study was to determine the association between increased INR and the usage of warfarin by using the pharmacovigilance data from the FDA Adverse Event Reporting System (FAERS).
Methods: The ADEs in patients who took warfarin was analyzed using FAERS data. Association rule mining was applied to identify warfarin-related ADEs that were most associated with elevated INR as well as possible drug-drug interactions (DDIs) associated with increased INR. In addition, this study sought to determine if the increased INR risk was influenced by sex, age, temporal distribution, and geographic distribution.
Results: The top 5 ADEs most associated with increased INR in patients taking warfarin were decreased hemoglobin, drug interactions, hematuria, asthenia, and fall. INR risk increased as age increased with individuals older than 80 having a 64% greater likelihood of elevated INR compared to those younger than 50. Males were 10% more likely to report increased INR as an ADE compared to females. Individuals taking warfarin concomitantly with at least one other drug were 43% more likely to report increased INR. The top 5 most frequently identified DDIs in patients taking warfarin and presenting with elevated INR were acetaminophen (lift = 1.81), ramipril (lift = 1.72), furosemide (lift = 1.64), bisoprolol (lift = 1.59), and simvastatin (lift = 1.59).
Conclusion: Patients prescribed warfarin were at greater risk of reporting increase INR. This effect may be less pronounced in women due to the procoagulatory effects of estrogen signaling. Multiple possible DDIs were identified including acetaminophen, ramipril, and furosemide.

Keywords: Warfarin, FAERS, FDA, pharmacovigilance, International Normalized Ratio, INR

Concurrent Session – Medical Informatics, Public Health Informatics and Pharmacoinformatics II
Tuesday, July 18, 2023
2:30 – 5:00 PM
Williams/Demens

Chairs: Eric Ho, Lijun Cheng

Predicting COVID-19 Severity of Emergency Room Patients using Chest X-ray Images

Jonathan Stubblefield[1,2,4,5], Christopher Saldivar[1,2,4], Anna De Feria[2,3], James Riddle[2,3], Abhijit Shivkumar[2,3], Jason Causey[1,2,4], Jake Qualls[1,2,4], Jennifer Fowler[1,2], Xiuzhen Huang[1,2,6]

[1] Center for No-Boundary Thinking (CNBT), Arkansas State University, Jonesboro, Arkansas

[2] The Joint Translational Research Lab of Arkansas State University and St. Bernards Medical Center, Jonesboro, Arkansas

[3] The Internal Medicine Residency Program, St. Bernards Medical Center, 225 E Jackson Ave, Jonesboro, Arkansas

[4] Department of Computer Science, Arkansas State University, Jonesboro, Arkansas

[5] Arkansas Biosciences Institute, Arkansas State University, Jonesboro, Arkansas

[6] Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, CA

Abstract.

In this study, we examined the severity of COVID-19 using a private dataset of chest x-ray images from 1,550 patients who tested positive for the virus and were admitted to the emergency room of St. Bernards Medical Center. Our investigation focused on two primary questions: Firstly, we sought to predict the length of hospital stay based on the chest x-ray images taken when patients were admitted to the ER. We found that predicting the duration of hospitalization using only chest x-ray images was challenging. None of the models we tested were better than a most-frequent classifier. However, when the data was split into four categories, each model outperformed the most-frequent classifier. This suggests that there is signal in the images, and the performance may improve with more data and clinical features. Secondly, we attempted to predict whether a patient had COVID-19 or not using chest x-ray images. We also examined the generalizability of the models by testing their performance on images captured from different sites. Using both our private dataset and the COVIDx dataset, our models achieved a high accuracy of 95.9%. However, we found that the model's performance suffered when the number of training samples was significantly reduced in any class. In conclusion, our study indicates that chest x-ray images may contain useful information for predicting the severity of COVID-19 and whether a patient is infected or not. However, further work is needed to improve the performance of the models, such as incorporating additional clinical features and increasing the training set size.

Keywords. Emergency room (ER), COVID-19, Chest x-ray image, Deep learning model

Quantifying the Growth of Glioblastoma Tumors Using Multimodal MRI Brain Images

Anisha Das, Shengxian Ding, Rongjie Liu and Chao Huang *
Department of Statistics, Florida State University, Tallahassee, Florida, USA
* Correspondence: chuang7@fsu.edu

Abstract:
Predicting the eventual volume of malignant or benign cells that might proliferate from a given tumor, can help in its early detection and subsequently the desired medical procedures can be applied to stop the proliferation of such cells thus preventing their migration to other organs. In this
work, a new formulation of the detection problem has been done using Bayesian technique for finding the eventual volume of such cells expected to proliferate from a Glioblastoma (GBM) tumor. The location of the tumor has been determined using parallel image segmentation algorithm. Once the location is determined, we find out how many cells can proliferate from this tumor until its survival time. For this, we start off by finding the likelihood of the eventual number of tumor cells that can take birth in a given time frame. We choose a certain prior subject to logic and our data set structure; and finally obtain our posterior distribution. Using the posterior mean, we obtain our desired eventual volume of tumor cells that need to be predicted. We also determine the corresponding probability that no tumor cell goes undetected when we find the ultimate eventual volume. The model so developed gives excellent results on our dataset. It is expected that the model will work on any dataset where the changes are not measured with regards time. We extend the model and run a Bayesian regression to incorporate other radiomic features of the tumor and discover that their inclusion enhances the chances of no tumor cells remaining undetected. We have mainly focused on detection of malignant cells, but the same model can be used for detecting both malignant and benign cells.

Keywords: Glioblastoma (GBM), malignant cells, proliferation, Bayesian technique, posterior mean

---

Comparing the risk of deep vein thrombosis of two combined oral contraceptives: norethindrone/ethinyl estradiol and drospirenone/ethinyl estradiol

Jennifer Stalas [1,$], Robert Morris [1,2,$], Kun Bu [3,$], Kevin von Bargen [1,$], Rebekah Largmann [1], Kathryn Sanford [1], Jacob Vandeventer [1], Weiru Han [2], and Feng Cheng [1,2,*]
[1] Taneja College of Pharmacy, University of South Florida, 12901 Bruce B Downs Blvd Tampa, 33612, USA
[2] Department of Biostatistics and Epidemiology, College of Public Health, University of South Florida, Tampa, FL 33612, USA
[3] Department of Mathematics & Statistics, College of Art and Science, University of South Florida, Tampa, FL 33620, USA
[$] These authors contributed equally to this work.
[*] Correspondence should be addressed to Feng Cheng; fcheng1@usf.edu

Abstract:
Background: Deep vein thrombosis (DVT) has been reported as an adverse event for patients receiving combined oral contraceptives. Norethindrone/ethinyl estradiol (NET/EE) and drospirenone/ethinyl estradiol

(DRSP/EE) are two commonly prescribed combined hormonal oral contraceptive agents used in the United States, differing in their progestin component.
Objective: The purpose of this study was to determine the association between the progestin component of a combined oral contraceptive (COC) and the risk of DVT in patients taking oral contraceptives for birth control using data derived from the FDA Adverse Event Reporting System (FAERS).
Methods: The risk of DVT was compared between patients that had taken NET/EE with those that had taken the DRSP/EE COC formulation for birth control. In addition, age was assessed as a possible confounder and the outcome severity for those diagnosed with DVT were compared between the two groups. Finally, association rule mining was utilized in order to identify possible drug-drug interactions that result in elevated DVT risk.
Results: DVT was the fourth most commonly adverse event reported for patients taking drospirenone/ ethinyl estradiol accounting for 8430 cases (20.28% of reports) and the fourteenth most commonly reported adverse event for norethindrone/ ethinyl estradiol accounting for 290 cases (2.3% of reports). Age was found to be a significant confounder for users of DRSP/EE with regards to DVT risk across all age groups assessed: 20<Age<=30 (ROR = 1.34, 95% CI 1.23-1.46), 30<Age<=40 (ROR = 2.16, 95% CI 1.98-2.35), and Age>40 (ROR = 3.68, 95% CI 3.36-4.03) However, there was only a statistically significant elevated risk in patients over 40 years of age taking NET/EE (ROR = 2.1221, 95% CI 1.45-3.11). Patients that had taken DRSP/EE and the corticosteroid prednisone simultaneously had an 8-fold increase in DVT risk (ROR = 8.08, 95% CI 7.02-9.31) relative to individuals that had only taken DRSP/EE.
Conclusion: Based on this analysis, there is a higher risk of developing DVT when taking drospirenone/ ethinyl estradiol than when taking norethindrone/ ethinyl estradiol as hormonal contraception. In addition, a possibly significant drug-drug interaction between different COC formulations and prednisone were identified. This interaction may result in elevated DVT risk due to a synergistic impairment of fibrinolysis and a decrease in plasmin production.

An In-silico Study of Antisense Oligonucleotide Antibiotics
Erica S. Chena, Eric S. Ho[1,*]
[1]Department of Biology, Lafayette College, Easton, PA 18042, USA
*Corresponding author. Email: hoe@lafayette.edu

Abstract
Introduction: The rapid emergence of antibiotic-resistant bacteria directly contributes to a wave of untreatable infections. The lack of new drug development is an important driver of this crisis. Most antibiotics today are small molecular compounds that block vital processes in bacteria. To achieve such effects, the 3-dimensional structure of targeted bacterial proteins must be resolved through purification followed by X-ray crystallography or NMR. However, such a task is time-consuming and tedious. Thus, improvement in antibiotic development is imperative. The development of RNA-based therapeutics has catalyzed a new platform of antibiotics – antisense oligonucleotides (ASOs). These molecules are complements to their target mRNA and prevent translation upon hybridization. This study aims to develop a bioinformatics pipeline to identify potent ASO targets in three bacterial species (P. gingivalis, H. influenzae, and S. aureus).
Methods: We downloaded open reading frames of bacterial essential genes from the Database of Essential Genes (DEG). After filtering for specificity and accessibility, ASO candidates were ranked based on their

self-hybridization score, predicted melting temperature, secondary-structure-free regions, and position on the gene. Gene enrichment analysis was conducted on putative potent ASOs.
Results: In P. gingivalis, we found 1117 ASOs in 348 essential genes, which corresponds to 43.91% of sequence space. In H. influenzae, we found 847 ASOs in 191 essential genes, which corresponds to 10.54% of sequence space. In S. aureus, we found 7061 ASOs in 346 essential genes, which corresponds to 89.11% of sequence space. Critical biological processes associated with these genes include translation, regulation of cell shape, cell division, and peptidoglycan biosynthetic process.
Conclusions: The results demonstrate that our bioinformatics pipeline is capable of identifying unique and accessible ASO targets in key bacterial species. Furthermore, the bioinformatics pipeline can be generalized to identify putative ASO targets in other bacterial genomes.
Conclusions: The results demonstrate that our bioinformatics pipeline is capable of identifying unique and accessible ASO targets in key bacterial species. Furthermore, the bioinformatics pipeline can be generalized to identify putative ASO targets in other bacterial genomes.

Reducing the Data for Radiation Cancer Therapy Quality Assurance

Maryam Albuainin [1,3] Richard Shaw [1,2] Shuang (Sean) Luan [1]
[1] Department of Computer Science, University of New Mexico, Albuquerque, New bMexico
[2] Dept of Radiation Oncology, University of New Mexico, Albuquerque, New bMexico
[3] Department of Computer Science, Imam Abdulrahman Bin Faisal, Al Jubail, Saudi Arabia

Abstract
Radiation therapy is one of our most effective tools for combating cancers. Armed with the megavoltage x-rays generated from a medical linear accelerator, modern radiation therapy uses sophisticated optimization algorithms from a treatment planning system to calculate a therapeutic plan that delivers a lethal radiation dose to the targeted tumor while protecting the nearby normal tissues and critical structures. The implementation of modern radiation therapy starts with commissioning and quality assurance of linear accelerators and treatment planning systems. During this process, extensive machine measurements are obtained and then imported into the treatment planning system to establish an accurate model of the accelerator. After this process, the system is ready for clinical use. This is a lengthy and tedious process, where thousands of data points are collected manually by large mechanical devices. In this research, we applied 3 data compression algorithms wavelet, Monte Carlo random sampling, and bottleneck shortest paths to radiation therapy physics data. We found that over 92% of the data we collected were redundant! We believe new commissioning protocols need to be developed, which will lead to significant workload reductions of clinical staff.

Keywords: Data compression, medical physics, quality assurance, radiation cancer therapy, random sampling, shortest path.

Exploring How Healthcare Organizations Use Twitter: A Discourse Analysis Anonymous Submission

Aditya Singhal and Vijay Mago

Lakehead University, Thunder Bay, Ontario, Canada

Abstract:
The use of Twitter by healthcare organizations is an effective means of disseminating medical information to the public. However, the content of tweets can be influenced by various factors, such as health emergencies and medical breakthroughs. In this study, we conducted a discourse analysis to better understand how public and private healthcare organizations use Twitter and the factors that influence the content of their tweets. Data was collected from the Twitter accounts of five pharmaceutical companies, two US and two Canadian public health agencies, and the World Health Organization. The study applied topic modeling and association rule mining to identify text patterns that influence the content of tweets across different Twitter accounts. The findings revealed that building a reputation on Twitter goes beyond just evaluating the popularity of a tweet in the online sphere. Topic modeling, when applied synchronously with hashtag and tagging analysis can provide an increase in tweet popularity. Additionally, the study showed differences in language use and style across the Twitter accounts' categories and discussed how the impact of popular association rules could translate to significantly more user engagement. Overall, the results of this study provide insights into natural language processing for health literacy and present a way for organizations to structure their future content to ensure maximum public engagement.

Keywords: Twitter; causality inference; association rule mining; healthcare organizations; topic modeling

---

The Association Between Bradycardia and the Use of Remdesivir

Gibret Umeukeje[1,$], Robert Morris[1,$], Weiru Han[2,$], Kun Bu[2], Jin Wei[3], Ruisheng Liu[3] and Feng Cheng[1,4,*]
[1] Department of Pharmaceutical Sciences, Taneja College of Pharmacy, University of South Florida, Tampa, FL 33613, USA
[2] Department of Mathematics & Statistics, College of Art and Science, University of South Florida, Tampa, FL 33620, USA
[3] Department of Molecular Pharmacology & Physiology, University of South Florida, Morsani College of Medicine, Tampa, FL 33613, USA
[4] Department of Biostatistics and Epidemiology, College of Public Health, University of South Florida, Tampa, FL 33613, USA
[$] These authors contributed equally to this work.
[*] Correspondence should be addressed to Feng Cheng; fcheng1@usf.edu

Abstract:
Background: Bradycardia has been reported as an adverse event for patients receiving remdesivir for the treatment of COVID-19.
Objective: The purpose of this study was to elucidate the association between bradycardia and the usage of remdesivir by using the pharmacovigilance data derived from the FDA Adverse Event Reporting System (FAERS).
Methods: The risk of bradycardia in COVID-19 patients who took remdesivir for treatment was compared with those of patients who took other approved COVID-19 treatments. In addition, this study sought to determine if the dysphagia risk was influenced by sex and dosage as well as if possible drug-drug interactions between remdesivir and other drugs in COVID-19 patients influenced the risk of subsequent bradycardia diagnosis.

Results: When compared to COVID-19 patients treated with REGEN-COV antibodies, dexamethasone, or Paxlovid, patients treated with remdesivir were approximately 4 times as likely to report bradycardia as an adverse drug event to FAERS than COVID-19 patients treated with REGEN-COV (7.97% vs. 1.63%) or dexamethasone (7.97% vs. 1.53%). In addition, none of the 561 COVID-19 patients whose case indicated treatment with Paxlovid reported bradycardia as an adverse drug event. No statistically significant difference in bradycardia risk was detected between men and women treated with remdesivir for COVID-19. In addition, dosage was not found to be a significant confounder of the association between the likelihood of subsequent bradycardia development and the use of remdesivir for COVID-19 treatment. Finally, COVID-19 patients co-prescribed tocilizumab, tamsulosin, or levetiracetam along with remdesivir were 98% more likely (OR = 1.98, 95% CI 1.34-2.91), 143% more likely (OR = 2.43, 95% CI 1.41-4.20), and 308% more likely (OR = 4.08, 95% CI 2.10-7.92) respectively to report bradycardia as an adverse drug event.
Conclusion: COVID-19 patients treated with remdesivir were at greater risk of reporting bradycardia as an adverse event than patients prescribed any other approved treatment for COVID-19 infection. This increase in bradycardia occurrence may be attributed to both overstimulation of the vagus nerve and increased mitochondrial damage induced by the remdesivir metabolite.

Keywords: FAERS; Pharmacovigilance; FDA; Remdesivir; Bradycardia.

Concurrent Session – Machine Learning/Deep Learning in Biomedical Research I
Tuesday, July 18, 2023
2:50 – 5:00 PM
St. Petersburg II, III

Chairs: Ying Zhang, Qian Liu

Revealing the impact of genomic alterations on cancer cell signaling with an interpretable deep learning model

Shuangxia Ren, Jonathan D. Young[1*], Lujia Chen[2], Xinghua Lu[1,2*,]
[1] Intelligent Systems Program, School of Computing and Information, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[2] Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[*] Correspond to: jdy10@pitt.edu, xinghua@pitt.edu

Abstract
Cancer is a disease of aberrant cellular signaling resulting from somatic genomic alterations (SGAs). Heterogeneous SGA events in tumors lead to tumor-specific signaling system aberrations, determining a tumor's aggressiveness and response to therapy. We interpret the cancer signaling systems as a causal graphical model, where SGAs cause functional aberrations of signaling proteins; the impact of such aberrations is propagated in the system through signal transduction; often, the propagation of signals eventually leads to changed gene expression. To represent such a system, we developed a deep learning

model called a redundant input neural network (RINN) with a transparent redundant input architecture. We used an L1 regularized objective function to infer causal relationships between input, latent, and output variables. We hypothesized that training RINN on cancer omics data would map the functional impacts of genomic alterations to latent variables in a deep learning model, thereby revealing hierarchical causal relationships between variables perturbed by different genomic alterations. Importantly, RINN makes the latent variables partially interpretable by allowing direct connections between SGAs and internal latent variables and learning a mapping between them. We show that using SGAs as inputs, RINN can encode their impact on the signaling system and predict gene expression accurately when measured as the area under ROC curves. We show that RINN can discover the shared functional impact (similar embeddings) of SGAs that perturb a common signaling pathway, such as those perturbing the PI3K, Nrf2, and TGF pathways. We also show that RINN can discover known causal relationships in cellular signaling systems.

DeepCORE: An interpretable multi-view deep neural network model to detect co-operative regulatory elements

Pramod Bharadwaj Chandrashekar[1,2], Hai Chen[3,4], Matthew Lee[3], Navid Ahmadinejad[3,4], Li Liu[3,4,*]
[1] Waisman Center, University of Wisconsin-Madison, Madison, WI, 53705, USA
[2] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53076, USA
[3] College of Health Solutions, Arizona State University, Phoenix, AZ, United States
[4] Biodesign Institute, Arizona State University, Tempe, AZ, United States
[*] Corresponding author: E-mail: liliu@asu.edu

Abstract
Gene transcription is an essential process involved in all aspects of cellular functions with significant impact on biological traits and diseases. This process is tightly regulated by multiple elements that co-operate to jointly modulate the transcription levels of target genes. To decipher the complicated regulatory network, we present a novel multi-view attention-based deep neural network that models the relationship between genetic, epigenetic, and transcriptional patterns and identifies co-operative regulatory elements (COREs). We applied this new method, named DeepCORE, to predict transcriptomes in 25 different cell lines, which outperformed the state-of-the-art algorithms. Furthermore, DeepCORE translates the attention values embedded in the neural network into interpretable information, including locations of putative regulatory elements and their correlations, which collectively implies COREs. These COREs are significantly enriched with known promoters and enhancers. Novel regulatory elements discovered by DeepCORE showed epigenetic signatures consistent with the status of histone modification marks.

A novel interpretable k-hop graph attention network model of integrative omics data analysis to infer target-specific core signaling pathways

Ruoying Yuan[1,2,*], Jiarui Feng[1,2,*], Heming Zhang[1,*], Yixin Chen[2], Philip Payne[1], Fuhai Li[1,3]

[1] Institute for Informatics, Washington University School of Medicine, St Louis, MO, U.S.A;
[2] Computer Science & Engineering, Washington University in St Louis, St Louis, MO, USA;
[3] Department of Pediatrics, Washington University School of Medicine, St Louis, MO, USA

*co-first authors
Fuhai.Li@wustl.edu

Abstract
With the advent of sequencing technology, large-scale multi-omics data have been generated to understand the diversity and heterogeneity of genetic targets and associated complex signaling pathways at multiple levels in diseases, which are critical targets to guide the development of personalized precision medicine. However, it remains a challenging task to computationally mine a few essential targets and pathways from a large number of variables characterized by the multi-level multi-omics data. In this study, we proposed a novel interpretable k-hop graph attention network model, k-hop GAT, to integrate the multi-omics data to infer the essential targets and related signaling networks. We evaluated the proposed model using the multi-omics data, i.e., genetic mutation, copy number variation, methylation, gene expression data of 332 cancer lines; and the experimentally identified essential targets. The validation and comparison results indicated that the proposed model outperformed the GAT and graph convolutional network (GCN) models. Signaling network inference can help researchers to identify novel therapeutic targets and signaling pathways for precision drug and drug combination prediction.

Proformer-based Ensemble Learning for Gene Expression Prediction

Lucy Nwosu, Xiangfang Li, Seungchan Kim, Lijun Qian, Xishuang Dong

Prairie View A&M University, Prairie View, TX, USA

Abstract
Predicting gene expression in DREAM Challenges 2022 is beneficial for gaining insights into gene function, biolog-ical pathways, and genes involved in regulating development, cell behavior, and signaling. In this study, a novel ensemble model that integrates “proformer” and ensemble learning tech-niques is proposed. The approach begins with preprocessing a large dataset of gene sequences provided by the DREAM Challenges 2022. Multiple “proformers” are then constructed, each capable of independently predicting gene expression. The resulting gene expression predictions from the “proformers” are combined through averaging weighted summation of individual prediction from each “proformer” to produce final predictions. Experimental results demonstrated that the proposed model is able to effectively enhance the prediction performance through comprehensive evaluation with various metrics, and even outper-formed the winner in the DREAM Challenges 2022.

Index Terms: Gene Expression Prediction, Transformer, Ensemble Deep Learning, Weighted Summation

---

DeepDecon accurately estimates cancer cell fractions in bulk RNA-seq data

Jiawei Huang[1], Yuxuan Du[1], Andres Stucky[2], Jiang F. Zhong[2*] and Fengzhu Sun[1*]

[1] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, 90089, CA, USA.

[2] Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, 92350, CA, USA.
[*] Corresponding author(s). E-mail(s): jzhong@llu.edu; fsun@usc.edu;

Abstract
Understanding the cellular composition of a disease-related tissue is important in disease diagnosis, prognosis, and downstream treatment. The recent advances in single-cell RNA sequencing (scRNA-seq) technique and extensive high-quality datasets have allowed the measurement of gene expression profiles for individual cells and further make it possible to deconvolve cellular composition. Here, we present DeepDecon, a deep neural network model leveraging single-cell gene expression information to accurately predict the fraction of cancer cells in bulk tissues. DeepDecon is trained based on single-cell RNA sequencing data and is robust to experimental biases and noises. It provides a refining strategy, where the cancer cell fraction is iteratively estimated by models trained on data with cancer cell fractions that are similar to the previously estimated fraction. When applied to simulated and real cancer data, it outperforms existing state-of-the-art decomposition methods including Scaden, Bisque, MEAD, RNA-Sieve, MuSiC, and NNLS, considering both accuracy and robustness. The DeepDecon method is available at https://github.com/Jiawei-Huang/DeepDecon

Keywords: Deconvolution, scRNA-seq, Deep neural network, Cancer tissue

---

Accurate prediction of functional effect of single missense variants with deep learning

Houssemeddine Derbel[1], Zhongming Zhao[2], Qian Liu[1, 3*]

[1] Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, USA
[2] Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3] School of Life Sciences, College of Sciences, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, USA
[*] Correspondence: qian.liu@unlv.edu

Abstract:
Estimating functional effect of missense variants in protein is a fundamental problem in proteomics for clinical medicine and protein engineering. Although the dissection of natively occurring variants provides a list of deleterious variants, high-throughput deep mutational experiments are used to generate a comprehensive investigation of variants for a single protein. To enable mutational dissection on millions of proteins, computational approaches were proposed, but they were mainly based on hand-crafted evolutionary conservation with limited accuracy. As the transformer model ESM is developed, precise approaches are needed to be assessed with functional effect of protein variants on tens of high-throughput experimental data. We propose a novel deep learning model, named Rep2Mut-V2. Rep2Mut-V2 takes advantage of learned representation from transformer models and is able to generate superior prediction of 27 types of functional effect of protein variants. Evaluated on 38 protein datasets with 118,933 single missense variants, Rep2Mut-V2 can achieve an average Spearman's correlation coefficient of 0.7.

Rep2Mut-V2's performance is much higher than the performance generated by the two recently released approaches ESM and DeepSequence, whose averaged Spearman's correlation coefficient is 0.41 and 0.49 respectively. Also trained on limited data, Rep2Mut-V2 still outperforms ESM and DeepSequence, demonstrating that Rep2Mut-V2 can extend high-throughput experimental analysis for more protein variants to reduce experimental costs. In summary, Rep2Mut-V2 generates accurate prediction of functional effect of single missense variants of proteins, and can be used to assist the interpretation of variants in human disease studies.

Keywords: functional effect, deep learning, single missense variant, precise estimation, high-throughput experiments

Concurrent Session – Machine Learning/Deep Learning in Biomedical Research II
Wednesday, July 19, 2023
9:30 AM – 12:20 PM
Williams/Demens

Chairs: Xiao Fan, Yijie Wang

Metastatic cancer expression generator (MetGen): A generative contrastive learning framework for metastatic cancer generation

Zhentao Liu [1,2], Yu-Chiao Chiu [3,4], Yidong Chen [5,6,§], Yufei Huang[1,2,4,§]

[1] Department of Electrical and Computer, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[2] Cancer Virology Program, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[3] Cancer Therapeutics Program, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[4] Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA
[5] Greehey Children Cancer Research Institute, The University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA
[6] Department of Population Health Science, The University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA
[§] Corresponding author: Yufei Huang (yuh119@pitt.edu)

Abstract
Despite significant advances in tumor biology and clinical therapeutics, metastasis remains the primary cause of cancer-related deaths. While RNA-seq technology has been used extensively to study metastatic cancer characteristics, challenges persist in acquiring adequate transcriptomic data. To overcome this challenge, we propose MetGen, a generative contrastive learning based on deep learning model. MetGen generates synthetic metastatic cancer expression profiles using primary cancer and normal tissue expression data. Our results demonstrate that MetGen generates comparable samples to actual metastatic cancer

samples, and we discuss the learning mechanism of MetGen. Additionally, we demonstrate MetGen's interpretability using metastatic prostate cancer and metastatic breast cancer. MetGen has learned highly relevant signatures in cancer, tissue, and tumor microenvironment, such as immune response and the metastasis process, which potentially fosters a more comprehensive understanding of metastatic cancer biology. The development of MetGen represents a significant step toward the study of metastatic cancer biology by providing a generative model which identifies candidate therapeutic targets for the treatment of metastatic cancer.

Keywords: metastatic cancer, deep learning, contrastive learning, tumor microenvironment

---

Neural relational inference optimization to analyze enzyme allosteric interactions in regular enzymes

Shuang Wang[1], Yan Wang[1], Yi He[2], Xuhong Zhang[3], Weiwei Han[2], Juexin Wang[4,*]

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China.
[2] Key Laboratory for Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun, China.
[3] Department of Computer Science, Luddy school of informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47408, USA
[4] Department of BioHealth Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University Purdue University Indianapolis, Indianapolis, IN 46202, USA
[*] Corresponding Author: Juexin Wang (wangjuex@iu.edu)

Abstract
Neural relational inference in molecular dynamics (NRI-MD) is useful for inferring high-order non-linear allosteric interactions with molecular dynamics simulations in enzymes. However, the current NRI-MD model and implementations are computationally intensive and only feasible in inferences for proteins with up to 100 residues on a typical GPU. To address this limitation, we optimize NRI-MD in sampling strategies and implementations to reduce GPU memory usage. We introduce numerical sampling, probability sampling, and enzyme domain sampling to condense the input enzyme sequences and replace model parameters with reduced size and apply half-precision to variables. In comparative experiments, the proposed optimizations can reduce GPU memory usage by at least 50% without compromising the model's performance. These optimizations significantly improve the current implementation and enable the NRI-MD allosteric interaction analysis in regular enzymes with lower memory cost.

CCLHunter: an efficient toolkit for cancer cell line authentication

Congfan Bu[1,2,#], Xinchang Zheng[1,2,#], Jialin Mai[1,2,3], Zhi Nie[1,2,3], Jingyao Zeng[1,2], Qiheng Qian[1,2,3], Tianyi Xu[1,2,3], Yanling Sun[1,2,3], Yiming Bao[1,2,3,*], Jingfa Xiao[1,2,3,*]

[1] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China

[2] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[#] Joint First Authors
[*] Corresponding authors: baoym@big.ac.cn (B. Y.), xiaojingfa@big.ac.cn (X. J.)

Abstract
Cancer cell lines are essential components in cancer research, yet accurate authentication of these cell lines can be challenging, particularly for consanguineous cell lines with close genetic similarities. We introduce a new method called Cancer Cell Line Hunter (CCLHunter) to tackle this challenge. This approach utilizes the information of single nucleotide polymorphisms, expression profiles, and kindred topology to accurately authenticate 1,389 human cancer cell lines. CCLHunter can precisely and efficiently authenticate cell lines from consanguineous lineages, and those derived from other tissues of the same individual. Our evaluation results indicate that CCLHunter has a complete accuracy rate of 93.27%, with an accuracy of 89.28% even for consanguineous cell lines, outperforming existing methods. Additionally, we provide convenient access to CCLHunter through standalone software and a web server at https://ngdc.cncb.ac.cn/cclhunter.

Keywords: Cancer cell line, Cell line authentication, Web server

---

Seizure prediction based on deep learning driven by nonlinear dynamics

Xiaoyan Wei[1§], ZhangZhen[2], Yi Zhou[3]

[1] Data center, Guangzhou Women and Children's Medical Center, National Children's Medical Center for South Central Region, Guangzhou Medical University, No.9 Jinsui Road, Guangzhou 510623, China
[2] School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China;
[3] Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong Province, China
[§]Corresponding author: Xiaoyan Wei; Email: weixy35@mail2.sysu.edu.cn

Abstract:
Background: Accurate prediction of epileptic seizures will effectively improve the quality of life of patients, but to achieve accurate, fast and stable prediction, the focus is to identify the preictal period of epileptic seizures so as to train the model. However, the transition state from preictal to ictal period often depends on expert experience and lacks a unified and clear scientific method.
Method: In this study, based on the theory and method of nonlinear dynamics, combined with the theory of statistics, the nonlinear dynamics indicators are calculated to define the preictal period, and deep learning methods was used to learn the space-time characteristics of the transition period, and then establish a new comprehensive prediction model and early warning system for seizures, so as to improve the accuracy, efficiency and stability of epileptic seizure prediction.
Result: The time range of the preictal was finally determined to be 1 hour before the seizure.The average accuracy of seizure prediction based on deep learning driven by nonlinear dynamics was 96.58%, FNR was 6.98%, and FPR was 8.51%. **Conclusion:**This study combines the advantages of nonlinear dynamics and

deep convolution network for EEG data analysis to predict epileptic seizures, and support relevant innovative research on clinical seizures and daily real-time monitoring of patients.

Keywords: Epilepsy; Seizure prediction; nonlinear dynamics; reference range; Deep learning;

A Transformer-Based Deep Learning Approach for Fairly Predicting Post-Liver Transplant Risk Factors

Can Li [1], Xiaoqian Jiang [2] and Kai Zhang [2]

[1] School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX [2] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
Kai.Zhang.1@uth.tmc.edu

Abstract
Liver transplantation is a life-saving procedure for patients with end-stage liver disease. There are two main challenges in liver transplant: finding the best matching patient for a donor and ensuring transplant equity among different subpopulations. The current MELD scoring system evaluates a patient's mortality risk if not receiving an organ within 90 days. However, the donor-patient matching should also take into consideration post-transplant risk factors, such as cardiovascular disease, chronic rejection, etc., which are all common complications after transplant. Accurate prediction of these risk scores remains a significant challenge. In this study, we will use predictive models to solve the above challenge. We propose a deep learning framework model to predict multiple risk factors after a liver transplant. By formulating it as a multi-task learning problem, the proposed deep neural network was trained on this data to simultaneously predict the fivepost-transplant risks and achieve equally good performance by leveraging task balancing techniques. We also propose a novel fairness achieving algorithm and to ensure prediction fairness across different subpopulations. We used electronic health records of 160,360 liver transplant patients, including demographic information, clinical variables, and laboratory values, collected from the liver transplant records of the United States from 1987 to 2018. The performance of the model was evaluated using various performance metrics such as AUROC, AURPC, and accuracy. The results of our experiments demonstrate that the proposed multitask prediction model achieved high accuracy and good balance in predicting all five post-transplant risk factors, with a maximum accuracy discrepancy of only 2.7%. The fairness-achieving algorithm significantly reduced the fairness disparity compared to the baseline model.

Keywords: Fairness, liver transplantation, risk prediction

DRLCOMPLEX: Reconstruction of Protein Quaternary Structures Using Deep Reinforcement Learning

Elham Soltanikazemi[†], Raj S. Roy[†], Farhan Quadir, Nabin Giri, Alex Morehead, Jianlin Cheng[*]
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO , USA 65211
[†] Equal Contribution
[*] Corresponding Author: chengji@missouri.edu

ABSTRACT
Predicted inter-chain residue-residue contacts can be used to build the quaternary structure of pro-tein complexes from scratch. However, only a small number of methods have been developed to reconstruct protein quaternary structures using predicted inter-chain contacts. Here, we present an agent-based self-learning method based on deep reinforcement learning (DRLCOMPLEX) to build protein complex structures using inter-chain contacts as distance constraints. We rigorously tested DRLCOMPLEX on two standard datasets of homodimeric and heterodimeric protein complexes (i.e., the CASP-CAPRI homodimer and Std_32 heterodimer datasets) using both true and predicted interchain contacts as inputs. Utilizing true contacts as input, DRLCOMPLEX achieved high average TM-scores of 0.9895 and 0.987 and a low average interface RMSD (I_RMSD) of 0.2197 and 0.8976 on the two datasets, respectively. When predicted contacts are used, the method achieves TM-scores of 0.73 and 0.76 for homodimers and heterodimers, respectively. Our experiments find that the accuracy of reconstructed quaternary structures depends on the accuracy of the contact predictions. Compared to other optimization methods for reconstructing quaternary structures from inter-chain contacts, DRLCOMPLEX performs similar to an advanced gradient descent method and better than a Markov Chain Monte Carlo simulation method and a simulated annealing-based method, validat-ing the effectiveness of DRLCOMPLEX for quaternary reconstruction of protein complexes. The source code and the instruction to reproduce the results are open sourced and available on GitHub repository:https://github.com/jianlin-cheng/DRLComplex.

Keywords reinforcement learning · protein complex · quaternary structure

---

Pan-cancer drug response prediction through tumor decomposition by cancer cell lines

Yu-Ching Hsu[1,2,3,4], Yu-Chiao Chiu[5,6], Tzu-Pin Lu[3], Tzu-Hung Hsiao[7,*], Yidong Chen[4,8,*]

[1] Bioinformatics Program, Taiwan International Graduate Program, National Taiwan University, Taipei 115, Taiwan
[2] Bioinformatics Program, Institute of Statistical Science, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan
[3] Institute of Epidemiology and Preventive Medicine, Department of Public Health, College of Public Health, National Taiwan University, Taipei 100, Taiwan
[4] Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[5] Department of Medicine, School of Medicine, University of Pittsburgh, PA 15261, USA
[6] UPMC Hillman Cancer Center, University of Pittsburgh, PA 15232, USA
[7] Department of Medical Research, Taichung Veterans General Hospital, Taichung 40705, Taiwan
[8] Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA
* Correspondence: d93921032@gmail.com (T.H.); ChenY8@uthscsa.edu (Y.C.)

Abstract
Tumor heterogeneity can affect the outcome of cancer treatments, and characterization of tumor heterogeneity may aid in the development of personalized therapeutics. Currently, large-scale cancer drug

sensitivity data became available for the collection of cancer cell lines. However, the tasks that translate pharmacogenomics knowledge from in vitro cancer cell lines to tumors remain to be addressed. In the present study, we adopted a deep-learning-based model, Scaden, and trained by cancer cell lines, termed Scaden-CA, to decompose tumors into proportions of cancer-type specific cell lines. Then, we predicted drug responses of tumor samples by devising an algorithm that incorporates the proportions of cancer cell lines from tumor decomposition and the drug sensitivity data from the drug screening PRISM dataset. The results showed that the Scaden-CA model has excellent performance with concordance correlation coefficients of more than 0.9. When applying the Scaden-CA model to bulk CCLE RNA-Seq dataset, the average proportions of cell lines being decomposed are > 70% across most cancers, which validated the potential of applying the Scaden-CA model to real-world data. We applied the decomposition model to TCGA samples, we then evaluated drug response data along with sample mutation status, and also closely examined both previously reported and novel Cancer-Mutation-Drug relationships identified in our analysis results. In summary, our Scaden-CA model and drug prediction algorithm enabled a new approach of studying drug sensitivity of tumors through decomposition by cancer cell lines, and the results were consistent with previous findings and shed light on future directions of drug repurposing.

Keywords: Pan-cancer, Drug response, Decomposition, Deep learning

Concurrent Session – Computational Methods for Aging and Brain Research
Tuesday, July 18, 2023
9:50 AM – 11:45 AM
St. Petersburg II, III

Chairs: Shaolei Teng, Guogen Shan

Clustering Alzheimer's Disease Subtypes via Similarity Learning and Graph Diffusion

Tianyi Wei[1,I], Shu Yang[1,I], Davoud Ataee Tarzanagh[1,I], Jingxuan Bao[1], Jia Xu[1], Patryk Orzechowski[1,2], Joost B. Wagenaar[1], Qi Long[1], Li Shen[1,II], for the ADNI[III]
[1] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
[2] Department of Automatics and Robotics, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland
[I] Equal contribution by T. Wei, S. Yang and D. Ataee Tarzanagh.
[II] Correspondence to li.shen@pennmedicine.upenn.edu.
[III] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database

ABSTRACT
Alzheimer's disease (AD) is a complex neurodegenerative disorder that affects millions of people worldwide. Due to the heterogeneous nature of AD, its diagnosis and treatment pose critical challenges. Consequently, there is a growing research interest in identifying homogeneous AD subtypes that can assist in addressing these challenges in recent years. In this study, we aim to identify subtypes of AD that represent distinctive

clinical features and underlying pathology by utilizing unsupervised clustering with graph diffusion and similarity learning. We adopted SIMLR, a multi-kernel similarity learning framework, and graph diffusion to perform clustering on a group of 829 patients with AD and mild cognitive impairment (MCI, a prodromal stage of AD) based on their cortical thickness measurements extracted from magnetic resonance imaging (MRI) scans. Although the clustering approach we utilized has not been explored for the task of AD subtyping before, it demonstrated significantly better performance than several commonly used clustering methods. Specifically, we showed the power of graph diffusion in reducing the effects of noise in the subtype detection. Our results revealed five subtypes that differed remarkably in their biomarkers, cognitive status, and some other clinical features. To validate the resultant subtypes further, a genetic association study was carried out and successfully identified potential genetic underpinnings of different AD subtypes.

Keywords: AD subtyping, unsupervised clustering, similarity learning, graph diffusion, brain MRI

Machine Learning Analysis for Studying Aging-Associated Hearing Loss

Safa Shubbar*, Qiang Guan*, Jianxin Bao†, John W. Hawks‡
∗Department of Computer Science, Kent State University, Kent, OH USA
† Department of Anatomy and Neurobiology {sshubbar, qguan}@kent.edu, Northeast Ohio Medical University, Rootstown, OH, USA jbao@gatewaybiotechnology.com
‡ Kent State University, Kent, OH, USA jhawks@kent.edu

Abstract
Presbycusis, or age-related hearing loss (ARHL), is a condition marked by a gradual decline in auditory sensitiv-ity, involving the loss of sensory cells and central processing functions associated with aging. The key features of ARHL include difficulty in hearing high-pitched sounds, reduced ability to understand speech in noisy or echoey environments, trouble with detecting quick changes in speech, and impaired ability to locate the source of sounds. We did data-driven cluster analysis (k-means clustering) on dataset C, this dataset includes subjects with hearing loss audiograms and audiometric shape parameters (n = 733 and 15 features)), dataset A (hearing loss audiograms) and dataset B (Audiometric shape parameters) to verify that the cluster structure described for each dataset was reproducible.

Index Terms: Hearing Loss, Aging, Age related, Unsupervised Machine Learning.

Vagus nerve stimulation and blood pressure modulate neuronal activity in the periventricular cerebellum

Maria Alejandra Gonzalez-Gonzalez [1,2,3,*]
[1] The Grass Foundation, Marine Biological Laboratory, The Grass Laboratory, Woods Hole, MA.
[2] Department of Pediatrics-Neurology, Baylor College of Medicine, Houston, TX
[3] Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, TX
[*] Corresponding author

Abstract
The cerebellum, a long-standing brain region important for motor balance, has recently been recognized to participate in other functions including reward behaviors, emotion, and cardiovascular maintenance. The

ventral part of the cerebellum contacts the roof of the fourth ventricle and we recently demonstrated that this area contains new cell populations within the subependymal area. These cells locate in the cerebellar periventricular zone (CbPVZ) and are characterized by unique electrophysiological profiles and pharmacology. Given their location in close proximity of the fourth ventricle and in the vicinity of the autonomic brainstem nuclei that integrate peripheral autonomic function carried by the vagus nerve, we reasoned that these cells may be the key to explain cerebellar role in autonomic integration. Here, we tested the hypothesis that vagus nerve modulates the activity in the CbPVZ and that this area responds to changes in blood pressure. We used neuromodulation strategies to monitor the CbPVZ electrophysiological activity in response to vagus nerve stimulation in rats. We furthermore evaluated the effect of blood pressure changes on CbPVZ by inducing hypertension pharmacologically. Our data show that the CbPVZ is modulated by the vagus nerve. Furthermore, we identified a subset of cells that are sensitive to the increase in blood pressure and a second subset related to reduction in blood pressure. Overall, our data shed new light onto the CbPVZ autonomic function related to blood pressure regulation, opening new perspectives to elucidate autonomic roles of the cerebellum.

A Machine Learning Based Multiple Imputation Method for the Health and Aging Brain Study-Health Disparities

Fan Zhang[1,2+], Melissa Petersen[1,2], Leigh Johnson[1], James Hall[1], Raymond F Palmer[1], Sid E. O'Bryant[1]
[1] Institute for Translational Research; Department of Pharmacology & Neuroscience, University of North Texas Health Science Center, Fort Worth, TX, USA
[2] Department of Family Medicine, University of North Texas Health Science Center, Fort Worth, TX, USA
[3] Department of Family and Community Medicine, the University of Texas Health Science Center, San Antonio, TX, USA.
[+] Address correspondence to: Fan Zhang, Ph.D. fan.zhang@unthsc.edu

Abstract

Background: The Health and Aging Brain Study-Health Disparities (HABS-HD) project seeks to understand the biological, social and environmental factors that impact brain aging among diverse communities. A common issue for HABS-HD is missing data. It is impossible to achieve accurate machine learning (ML) if data contains missing values. Therefore, developing new imputation methodology has become an urgent task for HABS-HD. The three missing data assumptions: (1) Missing Completely at Random (MCAR), (2) Missing at Random (MAR), and (3) Missing Not at Random (MNAR), necessitates distinct imputation approaches for each mechanism of missingness. Several popular imputation methods, including listwise deletion, min, mean, predictive mean matching (PMM), classification and regression trees (CART), and missForest, may result in biased outcomes and reduced statistical power when applied to downstream analyses such as testing hypotheses related to clinical variables or utilizing machine learning to predict AD or MCI. Moreover, these commonly used imputation techniques can produce unreliable estimates of missing values if they do not account for the missingness mechanisms or if there is inconsistency between the imputation method and the missing data mechanism in HABS-HD.

Methods: Therefore, we proposed a three-step workflow to handle missing data in HABS-HD. First, we explored the missingness in HABS-HD. Then, we developed a Machine Learning based Multiple Imputation method (MLMI) for imputing missing values. We built four ML-based Imputation models

(Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGB) and lasso and elastic-net regularized Generalized Linear Model (GLMNET)) and adapted the four ML-based models to multiple imputation by simple averaging method. Lastly, we evaluated and 45 compared MLMI with other common methods.
Results: Our results showed that the three-step workflow worked well for handling 47 missing values in HABS-HD and the ML-based Multiple Imputation method 48 outperformed other common methods in terms of prediction performance and change 49 on distribution and correlation.
Conclusion: The choice of missing handling methodology has a significant impact on 51 the accompanying statistical analyses of HABS-HD. The conceptual three-step workflow 52 and the ML-based Multiple Imputation method performs well for our Alzheimer's disease 53 models. They can also be applied to other disease data analysis.

Key Words: Alzheimer's Disease, Blood Biomarkers, Machine Learning, Multiple Imputation, Missing Data

---

Structure-learning-based causal comorbidities mining from UK biobank: an exploratory study for Alzheimer's disease

Yiheng Pan[1,2,&], Pingjian Ding[1,&], Zhenxiang Gao[1], Rong Xu[1,*]

[1] Center for Artificial Intelligence in Drug Discovery, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA
[2] Department of Computer and Data Science, Case Western Reserve University, Cleveland, OH 44106, USA
[*] Corresponding author: Rong Xu (rxx@case.edu)
[&] These authors contributed equally: Yiheng Pan, Pingjian Ding.

Abstract
Background: Identification of causal comorbidities is crucial for understanding disease pathogenesis, and the availability of large-scale observational data from biomedical data has provided considerable opportunities. To filter out the co-occurring diseases that arise through confounders, structure learning presents the best estimation of the real-world network by considering the potential dependencies between the diseases. However, the challenges in applying structure causal learning on biomedical data is the sparseness nature of biomedical data.
Methods: This study used a score-based structure learning algorithm NOTEARS and introduced an observation indicator to deal with the sparseness characteristics of biomedical data and called it NOTEARS_alpha. After applying proposed algorithm in UK Biobank data, we used maximum absolute weight in adjacency matrix to rank the diseases directly related to Alzheimer's disease for exploration. Based on manually created list of known risk factors for Alzheimer's disease, we compared NOTEARS_alpha, NOTEARS, and statistical method relative risk for comorbidity mining.
Results: The highest AUPRC and AUROC were observed in NOTEARS_alpha (AUPRC: 0.286, AUROC: 0.764), followed by NOTEARS (AUPRC: 0.209, AUROC: 0.759) and relative risk (AUPRC: 0.190, AUROC: 0.745).

Conclusions: This exploratory study was the first to present the capability of NOTEARS algorithm of mining the causal diseases from large-scale UK Biobank data and added an observational indicator to address the sparseness problem in biomedical data. For causal comorbidities discovery of Alzheimer's disease, our updated algorithm NOTEARS_alpha outperformed original NOTEARS and relative risk.

Keywords: disease comorbidity mining; probabilistic graph modelling; structure learning; causal discovery; directed acyclic graph; Alzheimer's disease

Concurrent Session – Single Cell Omics Data Modeling and Analysis
Tuesday, July 18, 2023
2:50 PM – 5:00 PM
St. Petersburg I

Chairs: Qianqian Song, Guangyu Wang

Improving cellular phylogenies through integrated use of mutation order and optimality principles

Sayaka Miura[1,2+*], Tenzin Dolker[1,2+], Maxwell Sanderford[1,2], and Sudhir Kumar[1,2,3*]

[1] Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, USA
[2] Department of Biology, Temple University, Philadelphia, Pennsylvania, USA
[3] Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia
[+] Co-first authors
[*] Co-Corresponding author: E-mails: sayaka.miura@temple.edu and s.kumar@temple.edu

Abstract
The study of tumor evolution is being revolutionalized by single-cell sequencing technologies that surveys somatic variation of cancer cells. In these endeavors, reliable inference of evolutionary relationship of single cells is a key step. However, single-cell sequences contain errors and missing bases, which necessitate advancing standard molecular phylogenetics approaches. We have developed a computational approach that integratively applies standard phylogenetic optimality principles and patterns of co-occurrence of sequence variations to produce more expansive and accurate cellular phylogenies. We found the new approach to also perform well for CRISPR/Cas9 genome editing datasets, suggesting that it can be useful for a range of applications. We apply the new approach to empirical datasets to showcase its use for reconstructing recurrent mutations and mutational reversals as well as for phylodynamics analysis to infer metastatic cell migration events between tumors.

Keywords: single-cell sequencing, phylogeny, mutational history, tumor evolution, metastasis

---

Gradient boosting reveals spatially diverse cholesterol gene signatures in colon cancer

Xiuxiu Yang[1], Justin L Couetil[2], Debolina Chatterjee[1], Valerie D Ardon[2], Jie Zhang[2], Kun Huang[1,2,3], Travis S Johnson[1,3,4,*]

[1] Department of Biostatisticsand Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States
[2] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, United States
[3] Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis University School of Medicine, Indianapolis, IN, United States
[4] Indiana Biosciences Research Institute, Indianapolis, IN, United States

Abstract
Colon cancer (CC) is the second most common cause of cancer deaths and the fourth most prevalent cancer in the United States. Recently cholesterol metabolism has been identified as a potential therapeutic avenue due to its consistent association with tumor treatment effects and overall prognosis. We used publicly available CC gene expression data to evaluate the relationship between cholesterol metabolism and CC prognosis, spatial gene expression, and protein immunohistochemistry. We conducted differential gene analysis and KEGG pathway analysis on paired tumor and adjacent-normal samples, identifying that bile secretion pathways were significantly downregulated. Bile is an important component of cholesterol homeostasis. We applied the traditional Cox Proportional Hazard (CPH) model and machine learning (ML) methods, such as Lasso Regression (LR), Random Forests (RF), and the eXtreme Gradient Boosting (XGBoost), to build prognostic models using these cholesterol/bile-acid related genes. These models were compared with each other and other published models. We demonstrate that using cholesterol metabolism genes with XGBoost models improves stratification of CC patients into low and high-risk groups. Two genes, ADCY5 and SLC2A1, were consistently identified by our models as being significant prognostic features, we identified that these models show unique spatial expression distributions in CC tissues.

Keywords: colon cancer (CC), cholesterol, prognostic genes, machine learning (ML)

---

scDemultiplex: An iterative beta-binomial model-based method for accurate demultiplexing with hashtag oligos

Li-Ching Huang[1,2], Lindsey K Stolze[1,2], Alexander Gelbard[3], Yu Shyr[1,2], Qi Liu1,[2,#] Quanhu Sheng[1,2,#]

[1] Department of Biostatistics, [2] Center for Quantitative Sciences and [3] Department of Otolaryngology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[#] To whom correspondence should be addressed. qi.liu@vumc.org; quanhu.sheng@vumc.org

ABSTRACT
Single-cell sequencing have been widely used to characterize cellular heterogeneity. Sample multiplexing where multiple samples are pooled together for single-cell experiments, attracts wide attention due to its benefits of increasing capacity, reducing costs, and minimizing batch effects. To analyze multiplexed data, the first crucial step is to demultiplex, the process of assigning cells to individual samples.

Inaccurate demultiplexing will create false cell types and result in misleading characterization. We propose scDemultiplex, which models hashtag oligo (HTO) counts with beta-binomial distribution and uses an iterative strategy for further refinement. Compared with five existing demultiplexing approaches, scDemultiplex achieves the highest accuracy in identifying singlets and removing negatives and multiplets.

Decoding ecosystem heterogeneity and transcriptional regulation characteristics of multi-subtype renal cell carcinoma

Kailong Xug[1], Jie Liu[1], Heng Yang[1], Jiang Li[1], Lixin Ma[1,*], Gang Dou[2,*,1] and Wang Yang[1,*]

[1] State Key Laboratory of Biocatalysis and Enzyme Engineering, School of Life Sciences, Hubei University, Wuhan, China
[2] College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, 266590, China
[*] Correspondence: malixing@hubu.edu.cn, dougang521@sdust.edu.cn, yangwang@hubu.edu.cn

Abstract: Renal cell carcinoma is a complex disease with several subtypes, and the tumor microenvironment plays a crucial role in disease progression and treatment response. To characterize the renal cell carcinoma ecosystem and its single-cell atlas characteristics, we performed single-cell analysis on 51 532 cells derived from 18 samples including clear cell carcinoma of the kidney, chromophobic cell carcinoma of the kidney, and benign adjacent tissues. We found that different subtypes of renal cell carcinoma exhibited distinct compositions and proportions of various cells in the tumor microenvironment. Analysis between tumor and immune cells reveals the characteristics of the tumor ecosystem, which is related to immunosuppression and poor prognosis of renal cell carcinoma. High-frequency tumor-associated macrophages and effector T cells were found in pRCCs, and macrophages interacted most abundantly with other TME components in the tumor microenvironment. In addition, by performing SCENIC analysis of T cells and macrophages, we identified ZEB1, PRDM1, HDAC1, and IRF5 as potential therapeutic targets that were significantly correlated with the prognosis of renal cell carcinoma. Our findings deepen our understanding of the renal cell carcinoma ecosystem and highlight the importance of accurately classifying patients and developing precise medical protocols based on differences in the tumor microenvironment.

Keywords: renal cell carcinoma; scRNA-seq;Single Cell Atlas; tumor microenvironment; transcription factors; survival curve

---

Improving cell type identification at single-cell level

Mostafa Malmir[1], Jinyan Li[2], Anita Omo-Okhuasuy[1], Umar Jamil[1], Yidong Chen[3,4], Yu-Fang Jin[1*]

[1] Department of Electrical and Computer Engineering, the University of Texas at San Antonio, San Antonio, Texas 78249, USA
[2] Department of Management Science and Statistics, the University of Texas at San Antonio, San Antonio, Texas 78249, USA

[3] Greehey Children's Cancer Research Institute, The University of Texas Health San Antonio, San Antonio, Texas, 78229, USA
[4] Department of Population Health Sciences, The University of Texas Health San Antonio, San Antonio, Texas 78229, USA
* Correspondence: Yu-Fang Jin (yufang.jin@utsa.edu)

ABSTRACT
Significance: Cell typing is one of the core components in analyzing single-cell RNA sequencing (scRNAseq) data and is of great interest in understanding biological processes such as drug responses and disease progression. Current cell-typing algorithms often rely on clustering all cells using gene expression profiles with machine learning or deep learning approaches and then assigning cell types to each cluster with expert-accumulated marker genes for known cell types. This cluster-based cell typing results in gene signatures being assigned at the cell subpopulation level rather than at single-cell resolution, foiling the advantage of single-cell technology, and specifically, unable to identify rare cells or sub-phenotypes or unknown cells. To address this limitation, we propose a new non-clustering statistical method that directly assigns cell types to individual cells. **Method and Results:** The method adopted multiple correspondence analysis (MCA) to calculate the distance of each gene to a cell in Barycentric coordinates. Since marker genes are highly expressed in a primary cell type with low expression in other cell types, we also
calculated the gene-cell affinity by evaluating expression levels and the number of marker genes for known cell types. Cell type association was determined by combining the gene-cell distance and affinity to assign cell type to each cell. Hyperparameters for the cell typing method were determined by optimizing the accuracy of cell typing for benchmark data. The proposed method was applied to 3 benchmark datasets and demonstrated an improved cell-typing accuracy compared with 6 existing methods including Cell-ID, AuCell, scCATCH, SCINA, and scSorter. **Conclusion**: Overall, our approach combines several innovative techniques, such as gene signature selection with gene-cell distance, marker gene evaluation, cell type association scoring, and gene marker weighting, to achieve highly accurate cell typing results. The proposed method has the potential to be a valuable tool for accurate cell typing using scRNAseq data and provides new insight into cell typing and the interpretation of cell typing algorithms with biological meaning.

Keywords: Single-cell RNA sequencing, Cell typing, Multiple correspondence analysis

---

Osteogenic Differentiation Potential of Mesenchymal Stem Cells using Single Cell Multiomic Analysis

Duojiao Chen[1], Xiaona Chu[1], Hongyu Gao[1], Patrick McGuire[1], Xuhong Yu[1], Xiaoling Xuei[1], Yichen Liu[1], Sheng Liu[1], Jill Reiter[1], Jun Wan[1], Yunlong Liu[1] and Yue Wang[1*]

[1] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana
* Correspondence: yuewang@iu.edu

Abstract:
Mesenchymal stem cells (MSC) are multipotent stem cells that can differentiate into multiple cell

types, including osteoblasts, chondrocytes and adipocytes. Osteoblast differentiation is reduced during osteoporosis development, resulting in reduced bone formation. Further, MSC isolated from different donors possess distinct osteogenic capacity. In this study, we used single-cell multiome analysis to profile the transcriptome and epigenome of MSCs from four healthy donors. Data were obtained from ~1,300 to 1,600 cells for each donor. These cells were clustered into four groups, indicating that MSC from different donors have distinct chromatin accessible regulatory elements for regulating gene expression. To investigate the mechanism by which MSC undergo osteogenic differentiation, we used the chromatin accessibility data from the single-cell multiome data to identify individual-specific enhancer-promoter pairs and evaluated the expression levels and activities of the transcriptional regulators. The MSC from four donors showed distinct differentiation potential into osteoblasts. MSC of donor one showed the largest average motif activities, indicating that MSC from donor one was most likely to differentiate into osteoblasts. The results of our validation experiments were consistent with the bioinformatics prediction. We also tested the enrichment of GWAS signals of several musculoskeletal disease traits in the patient-specific chromatin accessible regions identified in the single-cell multiome data, including osteoporosis, osteopenia, and osteoarthritis. We found that osteoarthritis-associated variants were only enriched in the regions identified from donor four. In contrast, osteoporosis and osteopenia variants were enriched in regions from donor one and least enriched in donor four. Since osteoporosis and osteopenia are related to the density of bone cells, the enrichment of variants from these traits should be correlated with the osteogenic potential of MSC. In summary, this study provides large-scale data to link regulatory elements with their target genes to study the regulatory relationships during the differentiation of mesenchymal stem cells and provide a deeper insight into the gene regulatory mechanism.

Keywords: Mesenchymal stem cells, single-cell multiome, osteogenic differentiation

---

Do Single-cell Hi-C Data Follow A Power Law Distribution?

Bin Zhao[1,2], Patrick Shen[3], and Lu Liu[2,*]

[1] Department of Statistics, North Dakota State University
[2] Department of Computer Science, North Dakota State University
[3] Davies High School
[*] Correspondence: lu.liu.2@ndsu.edu

Abstract:
Power law distributions are prevalent in many natural and social systems and can reveal underlying mechanisms and structures in complex systems. Single-cell Hi-C datasets have become increasingly popular in biology due to their ability to capture three-dimensional chromatin organization at the single-cell level. While bulk Hi-C data are known to follow the power law distribution, there has been no investigation of whether single-cell Hi-C data follow the same pattern. In this study, we analyzed power law distributions in single-cell Hi-C datasets ranging from base to 1Mb resolution, using two methods to test for power law behavior: hypothesis testing and likelihood ratio testing. Our results demonstrate that the majority of single-cell Hi-C data follows a power law distribution, indicating that power law behavior is a fundamental property of biological systems. Our findings can be applied to developing new computational methods for single-cell Hi-C data.

Concurrent Session – Cancer Informatics and Network Biology
Wednesday, July 19, 2023
9:50 AM – 12:00 PM
St. Petersburg II, III

Chairs: Noam Auslander, Xinna Zhang

CoMatch: a transfer learning model connecting in vivo finding to outcome prediction to distinguish prognostic/predictive biomarkers in breast cancer.

Abhishek Majumdar[1#], Aida Yazdanparast[1#], Huanmei Wu[2,3], Lang Li[1,2*] and Lijun Cheng[1*]

[1] Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, Ohio 43210
[2] Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, Indiana 46202
[3] Department of BioHealth Informatics, Indiana University, Indianapolis, Indiana 46202
[*] Corresponding author: Lang Li (lang.li@osumc.edu) and Lijun Cheng (lijun.cheng@osumc.edu)
Abhishek Majumdar, abhishek.majumdar@osumc.edu
Aida Yazdanparast, ayazdanp@iu.edu
Huanmei Wu, hw9@iupui.edu

Abstract

Background: The identification of biomarkers can support clinical decision-making and pave the way for both clinical and basic research scenarios in personalized medicine. A prognostic biomarker provides information about the patient's overall cancer outcome whilst a predictive biomarker provides information on the likelihood of response to a given therapeutic treatment. The large scale of drug screening in cancer cells would allow us to devise more effective strategies for clinical translation from research to clinical care than traditional clinical treatment. However, it has not yet been fully addressed how to identify shared prognostic and predictive biomarkers for patients' accordance of a large drug screening data of cell lines, which guide the development/use of tailored therapies from covariate for patients' stratification. **Method**: We present a transfer learning co-module matching model, called, **CoMatch** for prognostic/predictive biomarker identification. The unique CoMatch model bridges cancer cell sensitivity of drug response to clinical outcomes to identify patient populations who are sensitivity to the specific drug. This approach provides a pattern match to quantify the predictive and prognostic strength between cancer cells and tumors response to specific drug, in a self-consistent mathematical and biology network framework by integrating DNA copy number variation and mutation, and RNA gene expression profiles analysis. We use drug-screening studies on thousand cancer cells of multi-genome variation from Cancer Cell Line Encyclopedia (CCLE) and drug response from Cancer Therapeutics Response Portal (CTRP), and real patients' multi-genome variation and clinical outcome in TCGA. **Result**: The novelty co-module matching technology on multi-omics data is used to predict anticancer therapy benefit for ER-negative breast cancer medicine by gene expression profiles, mutation and copy number variation analysis. Four standard chemotherapy agents are simulated by their drug response on cancer cells and matching with tumors survival of ER− breast cancer patients treated with chemotherapies paclitaxel (**T**), doxorubicin (**A**), docetaxel (**D**) and cyclophosphamide

(**C**). The common patterns relationships between the drugs, multi-omics changes, and the phenotypes are detected systematically across cancer cells and patients. The similarity block of predictive biomarkers across gene expression, CNV and mutation are observed systematically. For instance, the common module markers of survival and doxorubicin consists of 21 genes, 12 genes from mRNA gene expression: *ZAP70, CLDN8, RCOR3, DUSP15, POU2AF1, CDKN2C, MAP7D2, SYK, KCNG3, ALG14, FN1* and *HIST1H2AC* and 9 genes from CVN: *FBXO6, MEGF6, MIR551A, PEX11A, KIF7, PER3, AP3S2, FBXO2* and *PTCHD2*. **Conclusion:** Our contribution is the data-driven transfer learning model, which naturally distinguishes the multi-omics prognostic versus predictive role of co-module biomarkers across cancer cells and patients. The research paved the way for personalized medicine and to further refine critical clinical decision system. This paper identified the dual roles of biomarkers in drug development. It sounds a cautionary note about the need to develop a stronger evidence base including robust in vivo validation prior to commercializing predictive and prognostic markers for cancer medicine. On the other hand, it provides molecular mechanism explanation to disease progression and drug resistance. R implementations of the suggested methods are available at https://github.com/abhishekmaj08/CoMatch

Keyword: Precision medicine, breast cancer, machine learning, cancer cells, drug screening
Availability and implementation

Identifying Significantly Perturbed Subnetworks in Cancer Using Multiple Protein-Protein Interaction Networks

Le Yang[1], Runpu Chen[1], Thomas Melendy[1], Steve Goodison[3], Yijun Sun[1,2,*]

[1] Department of Microbiology and Immunology [2] Department of Computer Science and Engineering University at Buffalo, The State University of New York, Buffalo, NY 14203 [3] Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, FL 32224
[*] Please address all correspondence to Dr. Yijun Sun (yijunsun@buffalo.edu).

Abstract
Detecting cancer driver genes and pathways is a core mission of some recent large-scale cancer genome studies. Network-based methods detect significantly perturbed subnetworks as putative cancer pathways by incorporating genomics data with the topological information of protein-protein interaction (PPI) networks. However, commonly used PPI networks have distinct topo-logical structures, making the results of the same method vary widely when applied to different networks. Furthermore, emerging context-specific PPI networks often have incomplete topological structures, which pose challenges for existing subnetwork detection algorithms. In this paper, we propose a novel method, referred to as MultiFDRnet, to address the above issues. The basic idea is to model a set of PPI networks as a multiplex network to preserve the topological structure of indi-vidual networks and meanwhile introduce dependencies among them and then detect significantly perturbed subnetworks on the modeled multiplex network to use all the structural information si-multaneously. To demonstrate the effectiveness of the proposed method, a large-scale benchmark study was performed on both simulation and cancer data. The experimental result showed that the proposed method is able to detect significant subnetworks jointly supported by multiple PPI networks and identify novel modular structures in context-specific PPI networks. The developed software and data are freely available at https://github.com/yangle293/MultiFDRnet.

Integrating and interpreting multi-omics data via novel k-hop graph neural network models to uncover core disease signaling pathways in medulloblastoma

Zitian Tang[1,2], Jiarui Feng[1,3], Yixin Chen[3], Philip Payne[1], Fuhai Li[1,4,#]

[1] Institute for Informatics (I2), Washington University, School of Medicine, St Louis, MO, U.S.A; [2] Division of Biological and Biomedical Sciences, Washington University, School of Medicine, St Louis, MO, U.S.A;
[3] Computer Science and Engineering, Washington University in St Louis, St Louis, MO, U.S.A; [4] Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, Missouri, United States.
[#] Correspondence Email: Fuhai.Li@wustl.edu

Abstract
Medulloblastoma (MB) is the most common malignant brain tumor that primarily affects infants and children. Four molecular subtypes of MB have been identified, namely WNT, SHH, Group 3 (G3), and Group 4 (G4). G3 and G4 MBs exhibit significantly worse clinical treatment outcomes and higher metastatic rates compared to WNT and SHH subtypes, and has shown a gender bias with more occurrences in male than female subjects. Multi-omics analysis characterizing cellular signaling cascades and pathways of MB samples have been generated, which provides a systematic, multi-level and holistic view of MB subtypes. Nevertheless, the molecular mechanism and core signaling pathways of G3 and G4 MBs remain unclear, and there is no well-identified gene biomarker associated with 3 phenotypes, i.e., poor survival rate, higher metastasis rate, and higher prevalence in males. In this study, we proposed a novel k-hop graph neural network (GNN) model to integrate and interpret the multi-omics data, and designed a visualization approach to annotate the signaling sources. We applied the model to rank the key biomarkers and uncover the core signaling pathways explaining the aforementioned three phenotypes of G3 and G4 MB subtypes. The model successfully identified a set of reported and novel biomarker genes. Interestingly, we identified a decrease in the expression of large tumor suppressor kinases (LATS) and a few other genes involved in the G2/M cell division checkpoint in only the male group, which may contribute to a higher prevalence of G3 and G4 MBs in males. The proposed model can be directly applied to other multi-omics data studies to rank key disease targets and infer core signaling pathways.

---

scGEM: unveiling the nested tree-structured gene co-expressing modules in single-cell transcriptome data

Han Zhang[1], Xinghua Lu[1,2], Binfeng Lu[3], Lujia Chen[1,#]

[1] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
[2] UPMC Hillman Cancer Center, Pittsburgh, PA
[3] Center for Discovery and Innovation, Hackensack Meridian Health, Nutley, NJ
[#] Corresponding Author: Lujia Chen, Ph.D.

Abstract
Single-cell transcriptome analysis has fundamentally changed biological research by allowing higher resolution of computational analysis at individual cell and subsets of cell types. However,
the majority of current methods concentrate on clustering cell subtypes based on whole transcriptome of cells, fail to fulfill the need of the recognition of the cellular programs that determine the specialization and differentiation of the cell types. In this study, we present scGEM,
a nested tree-structure non-parametric Bayesian model to reveal the gene expression modules (GEMs) reflecting transcriptome processes of in single cells. We show that scGEM can discover transcriptome processes shared among different types of cells as well as those reflecting highly specialized cell functions. We systematically examined the impact of sample size, model complexity, and pretraining on the performance of scGEM. We applied scGEM to triple-negative breast cancer single-cell RNAseq data and identified the underlying cellular programs of cells in the tumor microenvironments. Finally, we show that information acquired by scGEM can be used to deconvolute bulk RNA data to estimate the presence of cells belonging to a specific subtype and the state of transcriptomic programs within these cells. Altogether, we demonstrate that scGEM can be used to model the transcriptome programs of single cells, thereby unveiling the specialization and generalization of transcriptomic programs across different types of cells.

Keywords: single cell transcriptome, topic model, gene co-expressing module, nested tree structure, cellular program

---

Repurposing drugs for Group3 and Group4 medulloblastoma subtypes by inhibiting novel common core signaling targets

Fuhai Li[1,2,#], William Buchser[5], Clifford Luke[2], Di Huang[1], Ma. Xenia G. Ilagan[4], Joshua B. Rubin[2,5]

[1] Institute for Informatics (I2), [2] Department of Pediatrics, [3] Department of Genetics, [4] Department of Biochemistry and Molecular Biophysics, [5] Department of Neuroscience, Washington University School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA. [6] Tumor Initiation & Maintenance Program, NCI-Designated Cancer Center, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. [#] Email: Fuhai.Li@wustl.edu

Abstract
Medulloblastoma (MB) is one of the most malignant pediatric brain tumors. Four subtypes of MBs have been identified using integrative omics data analysis, i.e., Wnt, SHH, Group 3 (G3) and Group 4 (G4). Compared with Wnt and SHH subtypes, G3 and G4 MBs have unclear dysfunctional signaling pathways and have worse outcome. There is a lack of effective targeted therapies with less toxicity for G3 and G4 MBs. In this study, it was found that the G3 and G4 shared a large set of common up-regulated genes. Thus, it was hypothesized that G3 and G4 the core signaling pathways. To identify the common core signaling pathways and targets, the gene ontology and super-gene ontology analyses were conducted. In the results, the p53, MYC, HDAC1, CDKs the cell cycle related signaling pathways and other targets were found to be activated in G3 and G4 MBs. Consequently, a set of drugs, e.g., HDAC, mTOR, IKK, Aurora kinase, PI3K, JAK, ATPase (heart disease), MEK. as well as the CDKs, TUBB inhibitors were top-ranked to inhibit these activated signaling pathways by using the reverse gene signature and the uncovered common core signaling pathways. In conclusion, instead of identifying the subtype specific core signaling pathways, the novel

common core signaling pathways and targets in G3 and G4 MBs were investigated, and a set of targeted drugs were repurposed that can inhibit these active signaling targets as novel treatment regimens for G3 and G4 MBs.

Prediction of prognosis, immunotherapy and chemotherapy with an immune-related risk score model in endometrial cancer

Wei Wei[#1], Zhenting Huang[2#], Bo Ye[#1], Xiaoling Mu[3], Jing Qiao[4], Peng Zhao[5], Yuehang Jiang[6], Jingxian Wu[2,6,7*], Xiaohui Zhan[*1]

[1] Department of Bioinformatics, School of Basic Medical Sciences Chongqing Medical University, Chongqing, 400016, P. R. China
[2] Department of Pathology, School of Basic Medical Sciences, Chongqing Medical University, Chongqing, 400016, P. R. China
[3] Department of Gynecology, the First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, P. R. China
[4] Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[5] School of Basic Medical Sciences, Chongqing Medical University,Chongqing, 400016, P. R. China
[6] Department of Pathology, the First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, P. R. China
[7] Molecular Medicine Diagnostic and Testing Center, Chongqing Medical University, Chongqing, 400016, P. R. China
[#] Co-first authors：Weiwei, Zhenting Huang, Bo Ye
[*] Corresponding authors: xhzhan@cqmu.edu.cn; wujingxian@cqmu.edu.cn

Abstract
Endometrial cancer (EC) is the most common gynecologic cancer. The 44 overall survival remains unsatisfying due to lacking of effective treatment 45 screening approach. Therefore, there is strong need to develop an effective biomarker for prognosis and treatment response prediction. In this study, we employed co-expression network (GCN) analysis to mine immune-related GCN modules and key genes, and then an immune-related risk score model (IRSM) was constructed based on these immune-related key genes. IRSM was related to favorable prognosis and could play as an independent prognosis predictor. Meanwhile, the molecular basis and immune infiltrating cell contents of IRSM were revealed, and the functions of model genes in the carcinogenesis of EC were confirmed by IHC. After evaluating the associations of IRSM with clinical and molecular characteristics, for IRSM, the performance of prognosis prediction was validated and the ability to predict immunotherapy response was presented. Subsequently, we examined the relationship of IRSM with immunotherapy and chemotherapy separately. Patients in low-risk group were highly effect to immunotherapy and chemotherapy than in high-risk group. Interesting, patients responding to immunotherapy are also more sensitively to chemotherapy. Overall, we developed an effective model which could serve as a prognosis as well as treatment response predictor.

Key words: Endometrial Cancer, Immune-related key genes, Risk score model, Prognosis, Immunotherapy, Chemotherapy

Concurrent Session – Artificial Intelligence on Big Data: Promise for Early-stage Trainees
Monday, July 17, 2023
9:30 AM – 12:00 PM
Williams/Demens

Chairs: Yufang Jin, Chi Zhang, Yongsheng Bai

Abstract ID: 1260
Feasibility of a 3D Convolutional Neural Network for the Diagnosis of Alzheimer's Disease using Brain PET Scans

Troy Zhang[1],Yan Guo[2], and Yang Mi [2]
[1] Interlake High School, Bellevue WA 98008, USA [2] University of Miami, Miami FL 33136, USA

Abstract
Alzheimer's disease is a progressive neurodegenerative disease often characterized by a reduction in brain mass and deterioration of neural tissue. Conventional diagnosis of Alzheimer's typically includes a holistic review of a patient's medical history, psychological performance evaluations, and laboratory exams. Diagnoses made under this system, however, are often subjective and generally inaccurate, with overall diagnosis accuracy currently averaging 77%. Despite this, Alzheimer's frequently presents with elevated levels of beta amyloid plaques and abnormal tau proteins, which can be revealed using PET scans. As such scans typically contain faint patterns difficult for humans to discern, it is a fitting task for a deep learning and computer vision model. Such a model, although not substitutable for a conventional holistic review, could provide an objective score to enhance a physician's judgement. In this study, we investigate the feasibility of using a 3D convolutional neural network to provide a diagnosis for Alzheimer's. Using such a network, we attain a 3-fold cross-validated maximum validation accuracy of 90.36% and trial average performance of 89.49%, but with mean precision and recall scores of 51.35% and 50.48%, respectively. Ultimately, we find that such a network has the potential to be feasible in a clinical setting.

Keywords: Alzheimer's disease, convolutional neural network, positron emission tomography.

Abstract ID: 1025
Comparisons of Coronavirus Spike Proteins and the Mutation Effects on Virus-Host Interaction

Crystal Teng[1], Vidhyanand Mahase[2], Adebiyi Sobitan[2], Shaolei Teng[2]

[1]Northwest High School, Germantown, MD, 20874 [2]Department of Biology, Howard University, Washington, D.C., 20059 USA

Abstract
There are many different types of coronaviruses that have appeared. Some of the more dangerous ones towards people include SARS-CoV-2, SARS-CoV and MERS-CoV. During the infection process of COVID-19, the coronavirus spike (S) glycoproteins, which are seen as the "crowns" of the virus, progress the process by attaching to permissive cells and having their receptor-binding domains (RBDs) interact with human receptors on the host cell's membrane. Due to the importance of S RBDs and the role they play in the infection process of coronavirus, we decided to compare the S RBDs of MERS-CoV, SARS-CoV, SARS-CoV-2 original strain (Wuhan-Hu-1) and SARS-CoV-2 Omicron variants. We performed sequence alignments on the four S RBD sequences and generated the sequence percent identity for them. We also applied structure alignments to calculate the Root Mean Square Deviation (RMSD) values among the four S RBD structures. A method of computational saturation mutagenesis allowed for us to quantify the systemic effects of missense mutations on the protein-protein interaction. The results of these studies showed that specific residues, mainly 496, 498, 501, and 505, in the SARS-CoV-2 S RBD affected virus-host interaction. Differences in mutation effects on S binding between Wuhan-Hu-1 and Omicron variants were found as well. The findings help us understand Omicron's infection process and why it's so contagious. It provides areas we need to work on to make more effective vaccines and drugs to fight against COVID-19.

Keywords: COVID-19, spike glycoproteins, sequence alignment, structure alignment, computational saturation mutagenesis, virus-host interaction.

Abstract ID: 1002
Identification of Key Biomarkers Associated with Ductal Breast Cancer in Spatial Transcriptomics Data

Ellie Xi[1], Tutu Hu[2], Chloe Yu[3], Lulu Shang[4], Xiang Zhou[4], Allen Bai[5,6]

[1]BASIS Independent Silicon Valley, 1290 Parkmoor Ave, San Jose, CA 95126 USA;
[2]Tabor Academy, 66 Spring Street, Marion, MA 02738 USA;
[3]Union County Academy for Allied Health Sciences, 1776 Raritan Road, Scotch Plains, NJ 07076 USA;
[4]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 USA;
[5]Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197 USA;
[6]Next-Gen Intelligent Science Training, Ann Arbor, MI 48105 USA.

Abstract
Background: Spatial transcriptomics (ST) is a collection of groundbreaking genomic technologies that allows the measurement of gene expression with spatial localization information on tissues. Recent developments in ST analysis have allowed a deep investigation of breast cancer environments and the cellular composition of such malignancies. Despite these advancements, a comprehensive spatial exploration of breast cancer-related genes and their involvement in oncogenic signaling pathways has not yet been fully realized. This limitation is partly due to a lack of refined methods capable of effectively extracting differential gene expression information from multiple ST datasets. In this study, we integrated three 10x genomics breast cancer ST datasets that offer high-resolution gene expression insights. We identified potential breast cancer marker genes that are differentially expressed in tumor regions and our

study provided unique yet complementary information that can not be detected through traditional methods. **Method**: We downloaded the three, public breast cancer ST datasets from 10x genomics and applied SpatialPCA to identify spatial domains on each dataset. Then, we validated our SpatialPCAreported clusters with published pathologist annotations. We normalized each dataset with Scran's computeSumFactors and Seurat's SCTransform function separately to reduce bias and performed differential gene analysis using the Seurat R Package. The differential gene lists of the two normalization methods were then combined after gene set enrichment. Finally, we conducted survival, immune infiltration, and conserved domain analysis to report the biological functions of the detected candidates. **Result**: We identified 45 candidate genes found in the three ST datasets that potentially play significant roles in biological pathways underlying cancer maintenance and cell dysfunction. Out of the 45 genes, nine were identified as probable tumor suppressors in ductal carcinoma based on our survival analysis. These nine genes likely have significant roles in preventing cancer progression as antioncogenes and, therefore, have a high likelihood of serving as biomarkers for ductal breast cancer research. **Conclusion**: Investigating the interactions between abnormally expressed tumor genes allows for a deeper understanding of how breast cancer is instigated in the human body and, thus, gives greater insight into its therapies. In our study, we developed a pipeline of identifying cancer-risk genes from a multitude of ST datasets and elucidated their biological function in malignant breast tissues. We performed various downstream analyses such as survival analysis, protein-protein interactions, and conserved domain analysis to articulate the neoplastic tendencies of our identified genes. Our analysis would provide guidance for researchers to investigate potential therapeutic targets.

Keywords (up to six words)
Spatial transcriptomics, survival analysis, breast cancer, differential analysis, tumor suppressor

---

Assessing the Clinical Significance Identification Capability of DNA Language Models: A Study of Enformer's Performance on Disease-Causing Variants in Human Cis-Regulatory Elements

Rain Hou[1], Chang Li[1], Xiaoming Liu[1]

[1]USF Genomics and College of Public Health, University of South Florida, Tampa, FL, USA.

Abstract
DNA language models are specialized models designed to process and generate text sequences that represent DNA sequences. These models leverage the advancements in natural language processing techniques to understand and generate genetic sequences. Enformer is one of such DNA language models which was trained on human and mouse epigenomics profiles. It is illustrated to be the best performing language model in predicting gene expression levels. It also excelled in predicting the non-coding variants' effect on molecular phenotypes, such as eQTLs. However, no study has examined its ability in identifying and capturing relevant information in determining clinically significant variants. In this study, we will analyze disease-causing variants reported in genotype-phenotype databases (ClinVar and HGMD) in known human cis-regulatory elements, i.e. promoters and enhancers. We will systematically evaluate Enformer's ability in capturing relevant information from these variants and compare its performance with models specifically designed to predict these functional variants. Our study is expected to guide future application of such large-scale language models in clinical genetics settings.

Keywords: Language models, variants, ClinVar, eQTLs, HGMD, Enformer

---

Characterization of oncogenes and tumor suppressor genes with onco-microRNAs and tumor suppressor microRNAs

Claire Shen[1], Binze Li[2], Yongsheng Bai[3,4]

[1]Jordan High School, Fulshear, TX, USA; [2]Department of Statistics and Data Science, UCLA, Los Angeles, CA 90095; [3]Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197 USA; [4]Next-Gen Intelligent Science Training, Ann Arbor, MI 48105 USA

Abstract

Introduction: Kidney renal clear cell carcinoma (KIRC) is the most common subtype of kidney cancer. KIRC has a poor prognosis and a high mortality rate. To date, immunotherapy based on immune checkpoints is the most promising treatment. The Cancer Genome Atlas (TCGA) is an extensive cancer genomics program that aims to analyze human tumors and discover and catalog cancer-causing genome alterations. The TCGA has molecularly characterized over 20,000 primary cancers and matched normal samples across 33 types of cancer including KIRC. MicroRNAs (miRNAs) down-regulate target genes by binding to the 3' UTR of the respective targets. MiRNAs target genes at the post-transcriptional level after mRNA has been transcribed. The complexity of the miRNA-gene targeting interaction, as a result, has made this research more challenging. There are two classes of genes that this study focuses on which are oncogenes and tumor suppressor genes. Oncogenes function by deregulating cell proliferation and suppressing apoptosis. Tumor suppressor genes function by inhibiting cell proliferation and tumor development. There have been many studies that focus on the targeting relationship between genes and miRNAs, but few studies investigate the big picture of the relationship between oncogenes and tumor suppressor genes and how it can be regulated at the genome-wide level in the context of the miRNA targeting mechanism. **Methods and Results**: We downloaded the significant miRNA-gene targeting cluster files for KIRC from a published study. We then separated and annotated the tumor suppressors and oncogenes based on their survival analysis. We grouped the data into four different categories based on the miRNA-targeted gene relationship (TS-TS, Onco-Onco, TS-Onco, Onco-TS). There are 8 significant clusters for KIRC. Furthermore, the bioinformatics tool, TumorComparer, was utilized to obtain all KIRC genes within cell lines with a CNA rank greater than 0.5. We found that there was a greater number of genes with SNVs than CNAs for KIRC. Based on the survival significance (p-value < .05) analysis, we have identified 28 oncogenes and 188 tumor suppressor genes. We have also identified 8 tumor suppressor miRNAs and 62 onco-miRNAs. **Conclusions**: Our study reported more tumor suppressor genes than oncogenes and more onco-miRNAs than tumor suppressor miRNAs. These results confirmed the onco-miRNAs targeting TS genes ratio is similar to the TS miRNAs targeting oncogenes ratio and verified that our classification for TS and Onco categories is decent for both genes and miRNAs. Our study is valuable to medical researchers for cancer treatment.

Keywords: Cancer, KIRC, miRNA, Tumor Suppressor, Oncogenes, TCGA

Abstract ID: 1784

FLUXestimator: a webserver for predicting metabolic flux and variations using transcriptomics data.

Alex Lu[1,2], Zixuan Zhang[2+], Haiqi Zhu[2,3+], Pengtao Dang[2,4+], Jia Wang[2,3], Wennan Chang[2], Xiao Wang[2,3], Norah Alghamdi[2], Yong Zang[2,5], Wenzhuo Wu[6], Yijie Wang[3], Yu Zhang[2]*, Sha Cao[2,5]*, Chi Zhang[2]*

[1] Park Tudor School, Indianapolis, IN, US 46240
[2] Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, US 46202
[3] Department of Computer Sciences, Indiana University, Bloomington, IN, US 47405
[4] Department of Electric Computer Engineering, Purdue University, Indianapolis, IN, US 46202
[5] Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, US 46202
[6] Department of Industrial Engineering, Purdue University, West Lafayette, IN, US 47907

Abstract
Quantitative assessment of single cell fluxome is critical for understanding the metabolic heterogeneity in diseases. Unfortunately, laboratory-based single cell fluxomics is currently impractical, and the current computational tools for flux estimation are not designed for single cell-level prediction. Given the well-established link between transcriptomic and metabolomic profiles, leveraging single cell transcriptomics data to predict single cell fluxome is not only feasible but also an urgent task. In this study, we present FLUXestimator, an online platform for predicting metabolic fluxome and variations using single cell or general transcriptomics data of large sample-size. The FLUXestimator webserver implements a recently developed unsupervised approach called single cell flux estimation analysis (scFEA), which uses a new neural network architecture to estimate reaction rates from transcriptomics data. To the best of our knowledge, FLUXestimator is the first web-based tool dedicated to predicting cell-/sample-wise metabolic flux and metabolite variations using transcriptomics data of human, mouse and 15 other common experimental organisms. The FLUXestimator webserver is available at http://scFLUX.org/, and stand-alone tools for local use are available at https://github.com/changwn/scFEA. Our tool provides a new avenue for studying metabolic heterogeneity in diseases and has the potential to facilitate the development of new therapeutic strategies.

Keywords: Metabolism, Web server, Flux Analysis, Transcriptomics data

---

Abstract ID: 1350
Identifying relationships between cellular topology and gene expression in spatial transcriptomics of breast cancer tissues

Isabella Wu[1], Chen Li[2], Wentao Huang[2], Debolina Chatterjee[3], Jie Zhang[3], Chao Chen[2*], Travis S. Johnson[3*]

[1]Choate Rosemary Hall High School, Wallingford, Connecticut; [2]Stony Brook University, Stony Brook, New York; [3]Indiana University School of Medicine, Indianapolis, Indiana

*Corresponding authors: chao.chen.1@stonybrook.edu and johnstrs@iu.edu

Abstract
Breast cancer is one of the leading causes of cancer-related deaths worldwide. Like most solid tumors, pathology images provide useful information about the tumor microenvironment (TME) at diagnosis and computational pathology algorithms can quantify these images into useful features for diagnosis and prognosis. However, these techniques cannot account for all of the heterogeneity in the tissue especially that which is driven by underlying molecular changes. In tissues, imaging techniques reveal cell spatial organization, which is greatly affected by intercellular signaling and this signaling can be studied via transcriptome profiling. Spatial transcriptomics (ST) techniques enable the characterization of not only cell molecular information, but also their spatial organization. This facilitates the understanding of how gene expression is linked to spatial relationships and cellular topology allowing for the study of these functional connections in cancer. We propose a novel approach that directly connects topological features observed in cancer histology with gene expression data obtained through ST. Building upon our previous preliminary study, we applied topological data analysis (TDA) to the TME to extract 700 image topological features (ITFs) from breast cancer ST samples. Unlike our previous study, which solely identified genes correlated with these ITFs, we utilize correlation and significance analysis to construct the first comprehensive "Topo-Gene Dictionary" that establishes a direct link between all extracted ITFs and the expression of prominent genes. Using Topo-Gene Dictionary, the top 50 positively and negatively correlated genes for each ITF can be identified, and vice versa. ITFs strongly correlated with key immune signaling genes were selected as potential diagnostic, prognostic, or therapeutic markers. Functional enrichment analysis reveals that these ITF markers are significantly associated with key immune gene ontology terms, such as cytokine-mediated signaling and antigen presentation via MHC class I. This demonstrates that the ITFs may be potential representations of the interactions between tumor and immune cells, which is crucial to identify effective therapeutic targets. To validate the link between genes and topology, we assess the predictive power of the topological features in gene expression prediction models of key signaling genes. With Topo-Gene Dictionary, it becomes possible to explore patient cohorts even when only one type of data (gene expression or topology) is available, unlocking tremendous potential for mechanistic biological discovery, insights into how gene expression is linked to the spatial arrangement of cells, a more comprehensive understanding of the TME and cancer development.

Keywords: image analysis, topological data analysis, spatial transcriptomics, tumor microenvironment, breast cancer

---

Abstract ID: 1059
Pan-cancer analysis of metabolic shifts via flux estimation analysis
Kevin Hu[1,2], Alex Lu[1,3], Grace Yang[1,2], Shaoyang Huang[1,2], Pengtao Dang[1,4], Yijie Wang[1,5], Haiqi Zhu[1,5], Sha Cao[1,6], Chi Zhang[1,7]

[1]Center for Computational Biology and Bioinformatics, [6]Department of Biostatistics, [7]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA.
[2]Carmel High School, Carmel, IN, USA. [3]Park Tudor School, Indianapolis, IN, USA.
[4]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, USA.

[5]Department of Computer Science, Indiana University, Bloomington, IN, USA

Abstract
Glucose and glutamine are major carbon and energy sources that promote the Q9 rapid proliferation of cancer cells. Metabolic shifts observed on cell lines or mouse models may not reflect the general metabolic shifts in real human cancer tissue. In this study, we conducted a computational characterization of the flux distribution and variations of the central energy metabolism and key branches in a pan-cancer analysis, including the glycolytic pathway, production of lactate, tricarboxylic acid (TCA) cycle, nucleic acid synthesis, glutaminolysis, glutamate, glutamine, and glutathione metabolism, and amino acid synthesis, in 11 cancer subtypes and nine matched adjacent normal tissue types using TCGA transcriptomics data. Our analysis confirms the increased influx in glucose uptake and glycolysis and decreased upper part of the TCA cycle, i.e., the Warburg effect, in almost all the analyzed cancer. However, increased lactate production and the second half of the TCA cycle were only seen in certain cancer types. More interestingly, we failed to detect significantly altered glutaminolysis in cancer tissues compared to their adjacent normal tissues. A systems biology model of metabolic shifts through cancer and tissue types is further developed and analyzed. We observed that (1) normal tissues have distinct metabolic phenotypes; (2) cancer types have drastically different metabolic shifts compared to their adjacent normal controls; and (3) the different shifts in tissue-specific metabolic phenotypes result in a converged metabolic phenotype through cancer types and cancer progression. This study strongly suggests the possibility of having a unified framework for studies of cancer-inducing stressors, adaptive metabolic reprogramming, and cancerous behaviors.

Keywords: cancer metabolism, flux estimation, glutaminolysis, TCA cycle, system biology

---

Temporal Phenotyping for Transitional Disease Progress: an application to cardiovascular diseases and neurological diseases

Andy Wang[1,2], Cong Liu[2], Chunhua Weng[2]

[1]Peddie School, Hightstown, NJ, USA; [2]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA.

Abstract
Background: Patients suffering from complex diseases often experience the onset of comorbidities. One notable example is epilepsy, which is much more prominent in Alzheimer's (AD) patients than in the general population. Many other relationships have been discovered, including cardiovascular diseases (CVDs) and certain neurological diseases, but their correlation is less well-studied. Patients with one condition may transition into other seemingly unrelated illnesses, suggesting there exists underlying hidden relationships rather than random chance. Identifying diseases' transitional phenotypes makes it possible to predict and identify another condition's onset before symptoms occur. **Methods**: We utilized the Colombia Open Health Data (COHD) database which contains clinical concept prevalence and co-occurrence extracted from longitudinal electronic health records (EHRs) on over 5 million patients, including inpatient and outpatient data. Clinical concepts are standardized according to the OMOP Common Data Model (CDM). We used COHD's Python API for data analysis and visualization, where we analyzed temporal relationships among diseases. Additionally, the database provided methods to access potential correlations

between specific drugs or procedures and disease progression. As a benchmark for evaluating our findings and investigating the ability of EHRs to uncover novel insights, we employed ChatGPT 3.5 as a "control experiment" representing established human knowledge. This facilitated a comparative analysis and exploration of potential discoveries beyond the purview of human experts. **Results**: We first validated the findings of Kim et al (2020) on a different healthcare system, suggesting a possible causal relationship between epilepsy and the later onset of AD. We also identified phenotypic features within epilepsy subgroups that contribute to AD development. We expanded our analysis to investigate the influence of cardiovascular diseases on various later-onset neurological disorders and observed a substantial increase in the occurrence of Alzheimer's disease (AD) and dementia, with patients previously affected by myocardial infarction exhibiting a 6-fold and 8-fold higher risk, respectively. Although ChatGPT is able to identify the general correlation between neurological disorders and myocardial infarction, where it claims "the risk of dementia was approximately 35% higher in individuals who had experienced a myocardial infarction compared to those who had not", it has no knowledge on specific disease subtypes and how intermediate variables affect outcomes. **Conclusion**: Structured data from EHRs enable temporal phenotyping and the discovery of hidden temporal relationships among conditions, drugs, and procedures, but additional research is needed to validate these findings and untangle possible causal connections. We are extending this research to unstructured EHRs, focusing on rare diseases using the Open Annotation for Rare Diseases (OARD) database, which catalogs clinical phenotypes by standardized Human Phenotype Ontology (HPO) terms through natural language processing approaches.

Keywords: electronic health records, cardiovascular disease, neurological disease, temporal phenotyping, semantic relationship, machine-learning, OMOP, OHDSI

The artificial intelligence analysis of single-cell transcriptomes highlights the high heterogeneity in bladder cancer

Julia Chang[1,5,#], Xilin Wei[2,5,#], Canchen Chu[3,5,#], Alyssa Wang[4,5]

[1]Richard Montgomery High School, MD, USA, [2]Marriotts Ridge High School, MD, USA, [3]Grier School, PA, USA, [4]Winston Churchill High School, MD, USA, [5]White Oak Center for Gifted Youth

#: These authors contributed equally to this work

Abstract

According to The American Cancer Society, bladder cancer (BLCA) is the fourth most common cancer in men with estimates of 62,420 new cases and 12,160 deaths in the United States for 2023. In the US, despite the 5-year survival rate for BLCA being approximately 75%, it is only 4.6% survival for metastatic BLCA. The high mortality rate in BLCA might be due to high tumor heterogeneity that involves frequent DNA variations, cell sub-type diversity, and abnormal cell-cell communications. Previous studies on BLCA have been focused on the bulk level that only detects average whole genome expression of all cells without reflecting the cell complexity and diversity within tumors. Single-cell sequencing technologies have been emerging as powerful tools to investigate the heterogeneity of BLCA with ultra-high resolution and far more efficiently than bulk sequencing. Therefore, to explore tumor heterogeneity in BLCA, we conducted a comprehensive and systematic in-depth analysis of single-cell transcriptome data. We collected and

integrated 70,429 single-cell transcriptomes for BLCA from several public resources. After comparing different deep learning analytic tools, we selected "Seurat", a powerful tool installed in R with several artificial intelligence (AI) algorithms. All single-cell transcriptome data were processed by AI algorithms through the following steps: (i) Normalization and scaling by a centered log ratio transformation and a linear model; (ii) linear dimensionality reduction by PCA; (iii) Non-linear dimensional reduction by UMAP or t-SNE; (iv) clustering by SNN and Louvain algorithms; (v) Finding marker genes by the Wilcoxon Rank Sum test. The analysis results demonstrated that BLCA tumor could be classified into 15 major different clusters, indicating a very high intra-tumor heterogeneity of BLCA. We used cell-type specific genes (e.g., EPCAM, COL1A1, C1QB, PECAM1 and CD27) to annotate all clusters with 6 main cell subpopulations including epithelial cells, fibroblast, macrophage/myeloid cells, T cells, and B cells. A total of 4845 significant potential cancer marker genes (p-value<=0.01 and log2FC>=0.25) were identified. Visualization of the marker genes revealed significant specificity in their relative sub-populations. The differential analysis of marker genes between BLCA and normal bladder tissue demonstrated that COL1A1/ACTA2 and C1QB/CD86 were specific cancer markers for BLCA fibroblasts and macrophage/myeloid, respectively. In conclusion, the results highlighted high heterogeneity of BLCA and demonstrated the precision and effectiveness of AI algorithms in analyzing single-cell transcriptome data.

Tissue Domains Identification using Spatial Transcriptomics Data

Emily Wei[1], Karla Paniagua[2], David Andrew Cassel[3], Mario Flores[2], Yu-Fang Jin[2]

[1]Hillfield Strathallan College, Hamilton, Ontario, Canada; [2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, USA; [3]Department of Computer Science & Math, Arcadia University

Abstract

Spatial transcriptomics analysis has attracted significant amounts of research effort to acquire a deeper insight into the understanding of tissue organization and function, including a more comprehensive understanding of the relationship among gene expression levels, cell locations, and cellular morphology. Identification of spatial domains such as spatially distinct cell populations and differences in cell microenvironments is important to understand tissue functions and cellular architectures. Therefore, the goal of this study is to apply a reported algorithm, DeepST, to process 10× Visium spatial transcriptomics data for tissue domain identification with gene expression, cellular morphology, and the spatial location of cells. Tissues were first separated into small spots. The similarity of spots was determined with respect to the correlation of gene expression levels from all cells in a spot, similarity of cellular morphology in a spot, and distance among the spots. The morphological similarities of each spot were processed by a convolutional neural network and spatial coordinates were used to build an adjacency matrix to determine the neighborhood of spots. Multimodal vectors considering the similarity of spots were built for feature extraction and dimensionality reduction. The multimodal vectors were further processed by a deep learning algorithm to identify spatial domains. A public benchmark dataset, the human dorsolateral prefrontal cortex (DLPFC), was selected and processed using the DeepST algorithm. The dataset includes 12 slides from 151,676 samples and has been manually annotated with layers and white matter based on the morphological features and gene markers. Our results were consistent with the previous results that DeepST performed better than existing state-of-the-art methods.

Keywords: Spatial Transcriptomics, Deep Learning, Spatial Domains

---

Adaptive Deep Inference with Collaborative Architecture for IoT

Alejandro Villanueva[1], Yu-Fang Jin[1], Mimi Xie[2]

[1] Department of Electrical and Computer Engineering, University of Texas at San Antonio, USA;
[2]Department of Computer Science, University of Texas at San Antonio, USA.

Abstract
Deep Neural Network models (DNN) are being increasingly employed in IoT devices, enabling a new generation of smart applications. Due to the resource constraints of such edge devices, the DNN models deployed in them are of limited sizes. Although those models can accurately process most 'easy' inputs, they fail to process those 'difficult' inputs. The goal of this research is to study the existing approaches and find a way to optimize DNNs to strike a balance between increasing the accuracy and reducing the inference latency of the model while meeting the constraints. To achieve this goal, we explore novel methods to categorize inputs into multiple levels of difficulty such as easy, moderate, and hard levels, for typical applications such as image classification. We then examine an architecture that supports adaptive deep inference for a given input based on its difficulty level. Adaptive inference methods are utilized to correctly adjust the computational requirements of the architecture depending on the complexity of the input data. Specifically, we explore two dynamic approaches: the first solution performs selective computations with model partitioning that selectively skips some of the computational layers and blocks by inserting extra decision gates in the model. When certain classification confidence is higher than a set threshold, we can exit the model early. This allows adjusting the network depth according to the input scale; the second solution employs multiple DNN models in the IoT system and dynamically selects an appropriate one for processing each input at runtime. Both approaches can be implemented in resource-constrained environments that have limited processing and storage. The CIFAR-10 and MNIST datasets are used to train the DNN by providing it with images of a variety of objects with different classes.

Keywords: Adaptive Deep Inference, IoT, Accuracy, Latency

Analysis of thermal images for Nearby Animal Behavior using Deep learning architectures for enhancing vehicle safety

Eleni Avlonitis[1], Athanasios Ioannis Arvanitidis, Miltiadis Alamaniotis

Department of Electrical and Computer Engineering, University of Texas at San Antonio, USA

Abstract
Safety of vehicles is the most important concern in the automobile industry. Safety is attained in various stages and with a variety of means and its goal is the minimization of casualties and deaths observed due to vehicle accidents. One of the factors that cause a significant number of vehicle accidents every year is wild

animals Animal-car collisions is a frequently observed event because of the animal's unexpected behavior to near passing cars, and subsequently to the limited time win which the driver must react. For this reason, in this work we utilize deep learning, and more specifically, the YOLO architecture to analyze thermal images obtained by a camera mounted on a car. The goal of the analysis is to identify animal figures in the thermal images during nighttime. The difficulty in this analysis has to do with the extraction of the figure within complex backgrounds. Especially in summertime the radiated heat from the surroundings makes the detection and extraction of the animal figure a challenging task. In our work, we utilize deep learning on a real-world set obtained during nighttime in the San Antonio, TX areas. The obtained data contains thermal images of deer that stand nearby the passing car. Analysis results clearly designate the potential of the deep learning architecture – YOLO – to identify the deer figures. The overall goal of the work is to enhance the vehicle's safety by identifying and predicting the wild animal's behavior.

Keywords: Deep learning, animal detection, vehicle safety, thermal image analysis

Flash Talk Session
Sunday, July 16, 2023
2:00 – 4:45 PM
Williams/Demens

Chairs: Kaixiong Ye, Chengqi Wang

Abstract ID: 1003
Shared genetic basis informs the roles of polyunsaturated fatty acids in brain disorders

Huifang Xu[1], Yitang Sun[1], Michael Francis[2], Claire Cheng[1], Nitya Modulla[1], Kaixiong Ye[1,2]

[1]Department of Genetics, University of Georgia, Athens, Georgia, USA;
[2]Institute of Bioinformatics, University of Georgia, Athens, Georgia, USA;

Abstract
Polyunsaturated fatty acids (PUFAs) have been linked to brain disorders, and there is evidence suggesting that omega-3 supplementation may reduce the risks of major depression (MDD), anxiety, and anorexia. However, the effects of PUFAs on brain disorders remain inconclusive. Here, we systematically examine the shared genetic basis between the circulating levels of polyunsaturated fatty acids (PUFAs) and brain disorders, aiming to inform their causal relationships. We performed four major analyses using genome-wide association summary statistics for six circulating PUFA levels (N=114,999) and 20 brain disorders (N=9,725~762,917). The six PUFA traits include the percentages of total PUFAs, omega-3, omega-6, LA, and DHA in total fatty acids, and the omega-6 to omega-3 ratio. First, we performed genetic correlation (r_g) analysis with LD score regression (LDSC). We revealed a widespread and moderate genetic correlation between 16 brain disorders and most PUFA measures. All measures except the ratio have significant negative correlation with opioid dependence (r_g=-0.4~-0.23, $P < 0.05$), alcohol dependence (r_g=-0.3~-0.18, $P < 0.05$), and cannabis use disorders (r_g=-0.27~-0.20, $P < 0.001$). Moreover, these PUFA measures are positively correlated with obsessive-compulsive disorder (r_g=0.14~0.3, $P < 0.05$) and anorexia nervosa

(r_g=0.16~0.27, P < 2.27E-04). Second, we applied MiXeR to quantify the number of shared causal genetic variants. They ranged from 5 variants between omega-3% and Alzheimer's disease to 361 between PUFA% and MDD. Third, we applied Mendelian randomization (MR) analysis using TwoSampleMR to infer causal relationships. Our forward MR analysis identified nine pairs in which the PUFA measure may causally increase the brain disorder risk, and seven pairs in which PUFAs reduce the risk. Lastly, we performed colocalization analysis using HyPrColoc and statistical fine-mapping using SuSiE to identify colocalized regions and pinpoint shared causal variants. We revealed 40 colocalized regions (posterior probability > 0.7) shared among six PUFAs and ten brain disorders, 13 of which were unique. We pinpointed 22 unique, potential shared causal variants, such as rs1260326 (GCKR), rs174564 (FADS2), and rs4818766 (ADARB1). These findings highlight a widespread shared genetic basis between PUFAs and brain disorders, pinpoint specific shared genetic variants, and provide support for potential causal effects of PUFAs on brain disorders, particularly alcohol consumption, bipolar disorder, and major depression.

Keywords: Brain disorders, Polyunsaturated fatty acids, GWAS, Genetic correlation, Mendelian randomization; Colocalization;

Abstract ID: 1559

Common Genetic Variants are Associated with Plasma and Skin Carotenoid Metabolism in Ethinically Diverse US Populations

Yixing Han[1], Savannah Mwesigwa[2], Melissa N. Laska[3], Stephanie B. Jilcott Pitts[4], Nancy E. Moran[5], Neil A. Hanchard[1]

[1]Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD;
[2]Department of Medical Microbiology, College of Health Sciences, Makerere University, Kampala, Uganda;
[3]Division of Epidemiology & Community Health, School of Public Health, University of Minnesota,
[4]Department of Public Health, East Carolina University, Greenville, NC;
[5]USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX.

Abstract

Carotenoids, natural pigments found in plants, algae and some bacteria, possess antioxidant and anti-inflammatory properties and hold great potential in preventing various oxidative disorders, arteriosclerosis, obesity, and certain cancers. Blood and skin carotenoid concentrations serve as reliable biomarkers for assessing carotenoid intake from fruits and vegetables. However, the influence of genetic variability on carotenoid metabolism and its correlation with these biomarkers in diverse populations remains unexplored. Furthermore, the translation of findings from studies in European populations regarding the genetic control of metabolism to other populations remains unknown. We investigated the association between genetic variation and carotenoid levels in an ethnically diverse cohort comprising 207 adults from the United States, including individuals self-identifying as 'Black or African American' (59), 'Asian' (51), 'Non-Hispanic White or Caucasian' (70), and 'Hispanic' (27). Through comprehensive association analyses of genetic

variants and covariates, we assessed plasma carotenoid species and total skin carotenoid concentrations. Genome-wide genotyping was performed using the H3Africa microarray, complemented by targeted sequencing of 35 curated genes reported to be important in carotenoid metabolism. Principle Component Analysis against 1000 Genome Project Phase 3 dataset successfully clustered the 207 individuals and refined the self-reported race ethnicity. Genotyped SNPs were imputed with 1000 Genome Project and Haplotype Reference Consortium, resulting in 25,288,301 SNPs. After data quality control, 7,467,403 SNPs ($r^2 \geq 0.3$ and MAF $\geq 0.5\%$) were used for association analysis. Multivariate logistic regression analysis using PLINK and GEMMA identified novel genome-wide significant associations. Notably, variants in ATF6 (top SNP rs11579627) showed a significant association with plasma total carotenoids and lipid metabolism (p = 2.5x10e-8; Beta=0.34). Additionally, two variants in PKD1L2 (rs4889261 and rs7194871) were significantly associated with plasma beta-carotene. Pathway analysis using MAGMA linked the associated SNPs (p < 5x10e-6) to lipid metabolic pathways, aligning with the lipid soluble nature of carotenoids. Additionally, effect sizes by ancestry meta-analysis demonstrated that skin carotenoid metabolism is predominantly influenced by allele frequency differences across population ancestry clusters. Our research expands the current understanding of the genetic factors impacting carotenoid metabolism, unveiling a previously unrecognized role for variants involved in lipid metabolism. Morover, it emphasizes the importance of including individuals from diverse genetic ancestries in genomic studies of healthy metabolism. This implications of our research findings extend to health outcomes related to carotenoids and strengthens the understanding of skin and plasma carotenoid biomarkers for assessing dietary intake. Consequently, our study contributes to advancing the fields of precision nutrition and precision health.

Keywords Carotenoids, GWAS, genetic variants, racial ethnic populations, nutrition, biomarkers

---

Abstract ID: 1458

Cross-analysis between P. falciparum Var expression with host immunothrombosis markers to better define pediatric cerebral malaria phenotypes.

Iset Vera[1], Thomas Keller[1], Anne Kessler[2], Visopo Harawa[3,4,7], Wilson L. Mandala[3,4,5], Stephen J. Rogerson[6], Terrie E. Taylor[7,8], Karl B. Seydel[7,8], and Kami Kim[1]

[1] University of South Florida, Tampa, FL, USA
[2] Albert Einstein College of Medicine, Bronx, NY, USA
[3] Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi
[4] University of Malawi, College of Medicine, Biomedical Department, Blantyre, Malawi
[5] Academy of Medical Sciences, Malawi University of Science and Technology, Thyolo, Malawi
[6] The University of Melbourne, Melbourne, Australia
[7] Blantyre Malaria Project, Blantyre, Malawi
[8] Michigan State University, East Lansing, MI, USA

Abstract

Cerebral malaria (CM) is a severe manifestation attributable to infection with Plasmodium falciparum parasites. Cytoadhesion of infected erythrocytes to the brain microvasculature is mediated by the surface adhesive protein PfEMP1 encoded by var genes. Together with cytoadhesion, recruitment of immune cells (monocytes, platelets, and T-cells) leads to blood flow obstruction, inflammation, endothelial dysfunction

with blood brain barrier breakdown, and subsequent swelling and hypoxia. Clinical diagnostic methods such as retinopathy and magnetic resonance imaging to assess brain swelling have allowed for a better stratification of CM case definition. In this study we analyzed a comprehensive panel of soluble host immune markers in conjunction with parasite virulence factors (var gene expression) to identify a biomarker signature attributable to the CM phenotypes. We utilized machine learning approaches (Random Forest analysis, PCA and unsupervised analysis) to identify patterns distinguishing Ret- versus Ret+ cases of CM. The top 5 predictive features are immune factors associated with immunothrombosis - platelet count, Tissue Factor, Angiopoietin 2, host cell-free DNA, and MPO. Var gene cluster analysis distinguishes Ret- from Ret+ CM by the var A genes associated with EPCR binding (DC13), var B chimeric EPCR-ICAM1 binding PfEMP1 (DC8) and by the rosetting associated DC16 var. An additional parasite factor that distinguishes CM by retinopathy is cell-free P. falciparum DNA, pointing to a role in parasite DNA as a major antigen in promoting immune exacerbation and increased CM severity. Continued analysis will include predictive models for brain swelling as well as unsupervised approaches to identify alternate CM phenotypic groupings.

Keywords: Malaria, machine learning, precision medicine, var genes

---

Abstract ID: 1450

Unveiling Gene Interactions in Alzheimer's Disease by Integrating Genetic and Epigenetic Data with a Network-Based Approach

Keith Sanders[1], Astrid M Manuel[1], Andi Liu[1, 2], Boyan Leng[3], Xiangning Chen[1], Zhongming Zhao[1,2,4]

[1] Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX 77030, USA; [2] Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center, Houston, TX 77030, USA; [3] Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center, Houston, TX 77030, USA; [4] Human Genetics Center, School of Public Health, The University of Texas Health Science Center, Houston, TX 77030, USA

Abstract

Alzheimer's disease (AD) is a complex neurodegenerative disorder characterized by behavioral and cognitive impairments. Despite recent progress, the underlying etiology of AD remains elusive. Epigenetic mechanisms, particularly DNA methylation, present promising avenues for exploring gene-level regulation of AD development risks. This study aimed to integrate genetic, epigenetic, and protein-protein interactions (PPI) data using the dense module search of genome-wide association study (dmGWAS) tool to elucidate gene interactions associated with AD. Specifically, we utilized AD GWAS data from Wightman et al. (N = 1,126,563) and DNA methylation array data of human postmortem dorsolateral prefrontal cortex brain tissue sourced from ROSMAP (N = 551). Genetic gene-level Z-scores were calculated using the bioinformatics tool MAGMA. The Stouffer's Z-score method was applied to combine p-values of differentially methylated CpG sites at promoter regions, enabling the calculation of gene-level methylation Z-scores. The bioinformatics tool dmGWAS, utilized gene weights, methylation Z-scores, and a PPI sourced from the BioGRID PPI database to identify gene subnetworks associated with AD. Visualizations of the top-scoring modules were conducted with Cytoscape to reveal the finding of dmGWAS. Additionally,

gene set enrichment and drug target enrichment analyses were conducted to generate further insights. Our findings revealed 286 significant network modules, enriched with both GWAS and DNA methylation signals. The top module included nine genes with TRIM25 serving as the central node. Notably, BIN1 exhibited the largest node weight and was strongly hypomethylated. GNAS and EPHA1 were the most hypermethylated and hypomethylated genes, respectively. The subsequent evaluative enrichment analyses were performed with the top 10% of dmGWAS modules, comprised of 74 unique genes. Web-based Cell-type-Specific Enrichment Analysis (WebCSEA) highlighted highly significant enrichment of monocytes ($p < 5\times10^{-12}$). Functional enrichment analysis identified Gene Ontology (GO) terms significantly related to AD pathology (FDR < 0.05), including "amyloid-beta", "neurofibrillary tangle", and "tau protein binding", among others. Furthermore, drug target enrichment analysis using the Therapeutic Target Database identified 19 existing drug targets enriched in our gene networks (p-value = 0.03), including 5 FDA-approved drug targets, presenting potential therapeutic strategies for AD. Overall, we demonstrated a comprehensive approach, which gleaned insights into the gene interactions associated with AD by integration of genetic and epigenetic data. The findings of our study display genes both well-studied and underexplored in the context of AD pathogenesis. Moreover, our findings may provide supporting evidence for subsequent studies progressing our understanding of the disease.

Keywords: Alzheimer's disease (AD), Genome-wide association study (GWAS), epigenetics, DNA methylation, gene networks, systems biology

---

Abstract ID: 1353

MalariaSED: a deep learning framework to decipher the regulatory contributions of noncoding variants in malaria parasites

Chengqi Wang[1]*, Yibo Dong[1], Jenna Oberstaller[1], Chang Li[1], Min Zhang[1], Justin Gibbons[1], Camilla Valente Pires[1], Lei Zhu[4], Rays H.Y. Jiang[1], Kami Kim[2], Jun Miao[2], Thomas D. Otto[3], Liwang Cui[2], John H. Adams[1], Xiaoming Liu[1]

[1]Center for Global Health and Infectious Diseases Research and USF Genomics Program, College of Public Health, University of South Florida, Tampa, FL, USA.
[2]Department of Internal Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, USA
[3]School of Infection & Immunity, MVLS, University of Glasgow, Glasgow, UK.
[4]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

Abstract

Malaria remains one of the largest global public health challenges, with an estimated ~200 million cases worldwide in 2020. Epigenetic regulation plays a pivotal role in the control of diverse biological processes in malaria parasites, including antigenic variation, immune escape, invasion of red blood cells (RBCs) and the emergence of drug resistance. The contributions of noncoding variants to transcriptional regulation in malaria parasites remain elusive, mainly due to methodological difficulties in interpreting the functions of noncoding DNA. Sequence-based prediction models provide a new perspective on the chromatin effects of genomic variants. However, all currently available models were developed in model organisms, which is not suitable for malaria parasites due to several unusual features characterizing their genome architecture.

The most striking trait of the malaria parasite genome rendering existing tools inadequate is its extremely high AT content. Therefore, it is a top priority to develop a sequence-based prediction model specifically for malaria parasite chromatin profiles to identify and prioritize variants of concern to malaria control efforts. To address the challenge of identifying epigenetic regulatory effects of noncoding variants in malaria parasites, we developed a sequence-based, ab initio deep learning framework, MalariaSED, for 15 chromatin profiles in malaraia parasites. MalariaSED achieved high performance in predicting chromatin profiles in malaria parasites with an average auROC higher than 0.95, which is ~18% percent higher than the existing DL model previously developed in the mammalian genome (average auROC = 0.81 in Enformer). We further applied MalariaSED to successfully predict changes in PfApiAP2 TF-family binding profiles in response to CRISPR/Cas9-introduced mutations to their binding sites, demonstrating the utility of MalariaSED for assessing regulatory function of reported TF DNA-binding motifs in malaria parasites. The high predictive performance of MalariaSED allows us to carry out the first-ever regulatory functional annotation of ~1.3 million reported noncoding variants in P. falciparum parasites. Our analysis shows that the significant chromatin effects associated with geographically differentiated variants may lead to the perturbation of crucial biological processes related to parasite drug resistance (Artemisinin resistance, ART-R) and RBC invasion. The drug resistance associated with chromatin effects was further supported by the result when we applied MalariaSED to the eQTLs collected from a hotbed region of rising antimalarial resistance. ART-R mainly occurs during the early stage (ring stage parasite) of the parasite's erythrocytic cycle. The higher chromatin accessibility change is observed at the parasite ring stage surrounding reported eQTLs, which is in concordance with the skewness of these parasite cohorts.

Keywords: Deep learning model, malaria parasites, chromatin profile prediction

---

Abstract ID: 1177

Enhancing DNA Sequence Matching and Ranking through Deep Learning-Based Alignment-Free Model

Sumarga K. Sah Tyagi[1], Minh Pham[1], Yicheng Tu[1]

[1]Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

Abstract

DNA sequence matching and ranking play a crucial role in various tasks, such as variant calling, transcriptome assembly, and gene expression analysis. However, traditional sequence alignment methods face challenges in terms of scalability and accuracy, making them inadequate for handling the complexity of Next-Generation Sequencing (NGS) data. NGS generates massive amounts of genomic data at an unprecedented scale and speed. To address this limitation, it is imperative to develop more efficient and accurate methods for sequence matching and ranking. In order to overcome this bottleneck, we propose the utilization of a deep learning-based alignment-free model capable of learning and managing the intricacies of NGS data. This model aims to enhance the accuracy and scalability of sequence matching while ranking aligned sequences based on their similarity to the reference genome. Inspired by the success of neural networks in natural language processing, we present a framework based on a neural network architecture. Our framework employs recurrent nodes to construct representations from the sequences of bytes corresponding to the DNA sequences to be matched. These representations are then combined and fed into

feed-forward nodes, ultimately resulting in a prediction of sequence matching. To evaluate the performance of our proposed method, we conducted an extensive assessment using a substantial dataset comprising short reads (100 to 200bp), long reads (200 to 600bp), and very long reads (2k to 8k bp) collected from the National Center for Biotechnology Information (NCBI). The results of our evaluation demonstrate that the proposed framework can accurately predict sequence matches and effectively rank sequences from the entire knowledge base. This approach not only addresses the limitations of traditional sequence alignment methods but also leverages the power of deep learning to improve the efficiency and accuracy of DNA sequence matching and ranking.

Keywords: NGS, Deep Learning, Sequence prediction, ranking, Short reads, long reads

---

Abstract ID: 1526
3D genome reveals intratumor heterogeneity in Glioblastoma

Qixuan Wang[1], Juan Wang[1], Qiushi Jin[1], Mark W. Youngblood[1], Lena Ann Stasiak[1], Ye Hou[1], Yu Luan[1], Radhika Mathur[2], Joseph F. Costello[2], Feng Yue[1]

[1]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[2]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA

Abstract
Glioblastoma (GBM) represents the most prevalent malignant primary brain tumor, known for its highly unfavorable prognosis. Currently, most clinical biopsies are conducted at a single site, which may provide incomplete or misleading information regarding tumor characteristics and predicted response to therapy. While 3D genomics in adult gliomas have shed light on oncogenic mechanisms, the extent and significance of 3D genomic intra-tumoral heterogeneity (ITH) in GBM remain unexplored. To address these gaps, we performed Hi-C experiments in 21 samples obtained from 9 GBM patients, with 15 of them being spatially mapped based on their 3D coordinates. We showed that the most variable regions, both between patients and within the same patient, were situated in the inactive B compartment of the genomic region. We observed recurrent structural variation (SV) events and identified genes constantly implicated in SVs, such as CDKN2A/B, across different GBM patients. Our study uncovered extensive inter-tumoral and intra-tumoral heterogeneity at 3D genome level, encompassing A/B compartmentalization, chromatin interactions and structural variations. Notably, in a patient with 9 spatially mapped samples from both temporal and frontal regions, we successfully identified region-specific chromatin interactions, regulatory networks and key regulators within a single case. To our knowledge, this study represents the first large-scale exploration of the 3D genome in primary GBM patients and the initial investigation into the 3D genome within multiple regions of the same tumor tissue in GBM. Our findings provide unprecedented insights into the ITH of GBM at the 3D genomic level, opening new avenues for understanding and potentially targeting this devastating disease.

Keywords: Glioblastoma, 3D genomics, inter-tumoral and intra-tumoral heterogeneity, structural variation, regulatory networks

---

Abstract ID: 1743
Integrated Spatial Multi-omics Analysis Based on MALDI Data

Xin Ma[1,3], Cameron Shedlock[2,3], Harrison Clarke[2,3], Roberto Ribas[2,3], Terrymar Medina[2,3], Tara R. Hawkinson[2,3], Shannon Keohane[2,3], Craig W. Vander Kooi[2,3], Matthew S. Gentry[2,3], Li Chen[1,3], Ramon Sun[2,3]

[1]Department of Biostatistics, University of Florida, Gainesville, FL, USA;
[2]Department of Biochemistry and Molecular Biology, College of Medicine, University of Florida, Gainesville, FL, USA;
[3]Center for Advanced Spatial Biomolecule Research, University of Florida, Gainesville, FL, USA

Abstract
Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI-MSI) is a powerful technique used to investigate the spatial distribution of various molecules such as metabolites, lipids and glycans in tissues. Inspired by the correlation between multi-omics, here we introduce a novel workflow which integrates spatial multi-omics (metabolomics, lipidomics and glycomics) to segment the tissue into multiple clusters, detect the significant overrepresented peaks in clusters of interest, annotate significant peaks, and perform pathway enrichment analysis for clusters based on annotated peaks. To demonstrate the value of this workflow, we perform it on mouse brain tissue. After combining the multi-omics sets, the mouse brain tissue is divided into approximately 40 clusters, from which we select 10 clusters of interest, such as cortex, thalamus. By taking Wilcoxon test and calculating mean fold change between one selected cluster and all other reference clusters, we find top 50 significant overrepresented peaks for each cluster of interest and annotate part of them. The annotated peak list will be the input list for pathway enrichment analysis. Through pathway enrichment analysis, we could identify enriched metabolic pathways or functional categories associated with a set of significant compounds of interested clusters.

Keywords: MALDI, Spatial Multi-omics, Pathway Enrichment Analysis

---

Abstract ID: 1868
Multimodal machine learning combining image and textual data to predict rare genetic disorders

Da Wu[1], Jingye Yang[1], Kai Wang[1]

[1]Children's Hospital of Philadelphia, Philadelphia, PA, USA.

Abstract
Backgrounds: Rare diseases are individually rare, but collectively common: there are more than 10,000 known rare diseases that affect about 1 in 10 people (or 30 million people) in the US. Many of these disorders exhibit unique dysmorphic facial features that can provide valuable clues for recognizing a syndrome. Recent advancements in multimodal machine learning (MML), enabled by the Transformer architecture, offer promising opportunities to leverage different data modalities and enhance predictive capabilities. In light of these breakthroughs, we propose the utilization of ViLT, a cutting-edge

Transformer-based multimodal model capable of integrating both frontal facial images and clinical phenotypic text data from Electronic Health Records (EHR), to predict rare genetic disorders in patients. **Methods**: In our study, we conducted fine-tuning of the ViLT model using the GestaltMatcher Database (GMDB). This database comprises 7459 medical images, predominantly facial photos, of rare disorders obtained from publications and patients who provided appropriate consent through clinics. Additionally, the GMDB includes textual data containing Human Phenotype Ontology (HPO) terms. To enhance the training process, we expanded our training sets by retrieving texts and images from PubMed Central's open access individual article dataset. This expansion aimed to augment the robustness of the fine-tuned ViLT model. **Results:** By fine-tuning the model on the GMDB datasets, we were able to achieve an accuracy of ~80% in predicting relatively frequent disorders (those with more than 50 images available). We anticipate that our model's accuracy will further improve as we incorporate additional training sets, such as image-text pairs from PubMed Central's article dataset. This expectation is based on the well-known scaling property of Transformer-based models, which suggests that increased training data leads to enhanced performance. **Conclusion:** Our study demonstrates the significant potential of Transformer-based multimodal learning models in predicting rare genetic disorders through the integration of phenotypic textual and image data. Moving forward, we aim to extend the application of Transformer-based models to incorporate additional data modalities, such as video recordings, for the early detection of rare disorders. We believe that our multimodal biomedical AI (artificial intelligence) approach can be further generalized to address various other challenges in the biomedical domain, when multiple types of biomedical data modalities are available.

Keywords: Rare genetic disorders; Multimodal machine learning; Transformer

---

Abstract ID: 1770

A multimodal neuroimaging-based risk score for Alzheimer's disease by combining clinical and large N>37000 population data

Elaheh Zendehrouh[1,2], Mohammad SE. Sendi[2,3], Vince D. Calhoun[1,2]

[1] Department of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA, USA;
[2] Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA;
[3] McLean Hospital and Harvard Medical School, Boston, MA, USA

Abstract

Background: Alzheimer's disease (AD) is the most common type of dementia among people over 65. With no effective treatment available, a preventive approach is crucial. We aim to develop a new AD risk score based on brain imaging phenotypes, known as the brain-wide risk score (BRS). This research contributes to identifying individuals at high risk of developing AD and potentially enables targeted preventive measures. **Methods:** We utilized the OASIS-3 cohort, consisting of 1302 imaging samples, as a base dataset to generate control (CN) and mild cognitive impairment (MCI) groups. For the target population, we accessed cognitive scores and neuroimaging data from the UK Biobank study, including resting-state fMRI, sMRI, and demographic information of 37,784 individuals. Preprocessing was performed using SPM12,

and functional network connectivity (FNC) and gray matter (GM) were extracted. In the OASIS-3 dataset, we removed the mean of neuroimaging features and calculated the mean across CN and MCI groups. Similarly, in the UK Biobank dataset, we removed the mean of neuroimaging features and computed the distance between target data and CN/MCI reference data using correlation distance. Each participant in the UK Biobank dataset had two distance values, dist_CN and dist_MCI, which were used to calculate Δdiff = dist_CN - dist_MCI as an AD risk score (BRS). With two imaging modalities, we obtained two BRS values for each UK Biobank participant. Furthermore, we determined the 10th percentile of BRS for each modality and calculated the mean of neuroimaging features and their distance from the reference dataset within each percentile. **Results:** We found that there was a higher sensory FNC in the CN group compared to the MCI group. By utilizing the 10th percentile of the AD BRS estimated from FNC, we identified 10 distinct biotypes. Notably, lower percentiles corresponded to the MCI group, while higher percentiles closely resembled the CN group. Similarly, we identified 10 MCI biotypes based on GM. Again, lower percentiles represented the MCI group, while higher percentiles were more closely aligned with the CN group. **Conclusion:** In this study, we developed a new multimodal neuroimaging-based BRS. Based on the proposed BRS from each modality, we identified 10 MCI biotypes based on 10th percentile of BRS. In the next step, we would explore the link between neuroimaging-based BRS and cognition while exploring different distance metrics.

Keywords: functional network connectivity, mild cognitive impairment, multimodal markers, Alzheimer's disease biotype.

Abstract ID: 1429
Developing an Accurate and Interpretable Risk-Based Model for Lung Cancer Screening

Piyawan Conahan[1], Lary Robinson[2], Haley Tolbert[3], Margaret M Byrne[4], Lee Green[4], Yi Luo[1]

[1]Department of Machine Learning, Moffitt Cancer Center, FL, USA;
[2]Division of Thoracic Oncology (Surgery), Moffitt Cancer Center, FL, USA;
[3]Lung Cancer Screening Program, Moffitt Cancer Center, FL, USA;
[4]Department of Health Outcomes and Behavior, Moffitt Cancer Center, FL, USA.

Abstract
Background: Low-dose computed tomography (LDCT) can significantly reduce lung cancer mortality. However, eligibility for insurance coverage for LDCT is based on USPSTF guidelines that rely solely on age and smoking history. Therefore, current eligibility criteria may prevent coverage for, and thus screening of, substantial numbers of high-risk smokers. If we can more accurately pinpoint smokers at high risk, we can improve screening coverage and potentially enhance early detection. In this study, we identified additional characteristics and developed a Bayesian network lung cancer screening (BN-LCS) model that uses an explainable machine learning approach to predict lung cancer incidence and pinpoint smokers most in need of screening. These results were compared with those derived from using the 2021 USPSTF criteria. **Methods:** Our study utilized data from NCI's Prostate, Lung, Colorectal, and Ovarian (PLCO) screening trial, comprising 71,751 smokers enrolled between 1993 and 2001 with lung cancer diagnoses through 2009. Based on collection time, data was partitioned into 80% (57,401) for training and 20% (14,350) for validation. We first identified the most important features from 38 total features by integrating Markov

blanket and Random Forest approaches. We then constructed a BN for lung cancer prediction based on these features, classifying a smoker as LCS-eligible if their predicted risk exceeded a threshold. Performance evaluation employed 10-fold cross-validation, area under the receiver operating characteristic curve (AU-ROC), and 2000 stratified bootstrap replicates for AU-ROC confidence interval (CI) calculation. Finally, we compared BN-LCS model performance against the 2021 USPSTF criteria using McNemar's Test and the validation dataset. **Results:** Our BN-LCS model includes variables such as age, smoking history, emphysema status, personal and family cancer history, and X-ray history. The BN-LCS model yielded an AU-ROC of 0.77 (95% CI: 0.760–0.775) on training data and 0.76 on validation data. Compared to the 2021 USPSTF criteria, our BN-LCS model significantly improved sensitivity (79.73% vs. 76.01%, p=0.036) without loss of specificity (59.35% vs. 59.59%, p=0.557). It identified 22 additional lung cancer patients (3.72%) from 592 cases in the validation dataset, reducing missed detections by 15.49% versus USPSTF criteria. **Conclusions:** Our proposed BN-LCS model significantly outperforms 2021 USPSTF criteria in sensitivity based on numerical experiments. We can more accurately pinpoint those at high risk and could potentially enhance early detection. However, this model still needs validation with other independent datasets. Current USPSTF criteria should be a starting point for future refinements to identify additional features for high-risk groups needing screening.

Keywords
Lung cancer screening, Lung cancer risk prediction, Bayesian networks, Sensitivity analysis

---

Abstract ID: 1987
In silico Improvement of Highly Protective Antimalarial Antibodies

Mateo Reveiz[1], Andrew Schaub[1], Young Do Kwon[1], Prabhanshu Tripathi[1], Azza Idris[1,2], Amarendra Pegu[1], Lais Da Silva Pereira[1], Patience Kiyuka[1], Myungjin Lee[1], Tracy Liu[1], Chen-Hsiang Shen[1], Baoshan Zhang[1], Yongping Yang[1], Peter D. Kwong[1], Reda Rawi[1]

[1]Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA. [2] The Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA, 02139, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1352
Comprehensive Investigation of Active Learning Strategies for Anti-Cancer Drug Response Prediction

Priyanka Vasanthakumari[1], Yitan Zhu[1], Thomas Brettin[2], Alexander Partin[1], Maulik Shukla[1], and Rick L. Stevens[1,3]

[1] Division of Data Science and Learning, Argonne National Laboratory, Lemont, IL, USA
[2] Computing, Environment and Life Sciences Directorate, Argonne National Laboratory, Lemont IL, USA
[3] Department of Computer Science, The University of Chicago, Chicago, IL, USA

Abstract
It is well-known that cancers of the same histology type can respond differently to a treatment. Thus, computational drug response prediction is paramount for preclinical drug screening studies and clinical treatment design. To build drug response prediction models, treatment response data need to be generated through screening experiments and used as input to train the prediction models. In this study, we investigate various active learning strategies of selecting experiments to generate response data for (1) improving the performance of drug response prediction models built on the data and (2) identifying effective treatments. Active learning is an iterative model building approach whereby the data collected in each batch of experiments are used to improve the prediction model, which then aids in selecting the next set of experiments. Here, we focus on constructing drug-specific response prediction models for cancer cell lines. Various sampling approaches have been designed and applied to select cell lines for screening, including random, greedy, uncertainty, diversity, a combination of greedy and uncertainty, sampling-based hybrid approach, and iteration-based hybrid approach. All the sampling strategies except random are based on model predictions on candidate experiments and can be referred to as active learning approaches. In the sampling-based hybrid approach, some samples selected are from random sampling, and the active learning approach selects the rest. In the iteration-based hybrid approach, the first few iterations use random sampling while the rest uses active learning methods. We apply and evaluate these sampling approaches on an existing large-scale cell line drug screening dataset, the Cancer Therapeutics Response Portal v2 data. Prediction models are built using the LightGBM algorithm to predict the normalized area under the dose response curve (AUC) based on the gene expression profiles of cell lines. All the approaches are evaluated and compared using two criteria: 1) the number of identified hits that are defined as selected experiments validated to have an AUC < 0.5, and 2) drug response prediction performance measured by the root mean squared error or R-squared. The analysis was conducted for 100 drugs, and the results show a significant improvement in identifying hits with active learning than random sampling. Active learning methods have also demonstrated an improved response prediction performance for some drugs compared with random sampling.

Keywords: Computational drug response prediction, cancer, active learning, machine learning

---

Abstract ID: 1234

Bioinformatics and machine learning based identification of potential oxidative stress and glucose metabolism diagnostic Biomarkers in Alzheimer disease.

Sidra Aslam[1], Fatima Noor[2], Thomas G. Beach[1], Geidy E. Serrano[1]

[1]Banner Sun Health Research Institute, Sun City, AZ, USA
[2] Government College University Faisalabad, Pakistan

Abstract
Introduction: Alzheimer's disease (AD) is a devastated neurodegenerative disease, accounting for 60 to 80% of dementia cases. We do not fully understand AD etiology and pathogenesis but oxidative stress plays key roles in AD pathogenesis. Glucose metabolism is the main source of energy for brain and any trouble in its metabolism could lead to neuronal dysfunction. Many studies described the interplay between glucose

metabolism and oxidative stress in AD. We aimed to find an oxidative stress and glucose metabolism related gene (OSGMG) based diagnostic feature biomarkers for AD. **Method:** RNA seq data (GSE125583) generated from Brain and Body donation program (BBDP) cases is retrieved from GEO database (219 AD cases and 70 controls). Glucose metabolism and oxidative stress related genes are collected from MSigDb database. Limma package was used to identify the differentially expressed genes (DEGs) in AD which meet the following criterion: adj.P-value < 0.05. Differentially expressed oxidative stress and glucose metabolism related genes (DE-OSGMGs) were screened by the intersection of DEGs and OSGMGs in R. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses of DE-OSGMGs was run using R (clusterProfiler). Machine learning algorithms (SVM, WGCNA, and LASSO) were used to select the diagnostic feature biomarkers for AD. ROC curve was employed to assess the discriminatory power of diagnostic feature biomarkers. **Results:** Three hundred common genes between oxidative stress and glucose metabolism-related genes were identified as OSGMGs. While 1400 DEGs were found by analyzing RNA seq data (GSE125583). A total of 200 DE-OSGMGs were identified in this study, which were used for downstream analysis. KEGG analysis showed that the DE-OSGMGs were largely enriched in amyloid β degradation, glycolysis/gluconeogenesis and carbon metabolism, while UBD, PAK1, GSK3B, AKT1 were recognized as diagnostic feature biomarkers for AD via WGCNA, SVM, and LASSO algorithms**. Conclusion:** Our study identified and validated the molecular mechanisms of interplay of OSGMGs in AD, So, UBD, PAK1, GSK3B, AKT1could serve as potential diagnostic biomarkers for AD.

Keywords: Machine learning, Alzheimer disease, biomarker, Bioinformatics, oxidative stress, Glucose metabolism

---

Additional Flash Talks

Abstract ID: 1838
PROMPT BIOINFO. CASE STUDY: Shotgun Metagenomic Data Analysis

Zhu Xing[1,2], Qiyun Zhu[1,2*]

[1] School of Life Sciences, Arizona State University, Tempe, Arizona 85281, USA
[2] Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, Arizona 85281, USA
[*] Contact author: qiyun.zhu@asu.edu

Abstract:
Metagenomics is a pioneering approach that leverages the power of whole genome shotgun sequencing data to study microbiomes. This strategy exploits sequencing reads from heterogenous microbes within one sample, offering comprehensive understanding of microbiome complexity. The central objective of metagenomics is to precisely identify microbial composition present within given samples. The output is a feature table that describes the taxonomy and abundance of microbes by sample, ready for downstream analyses to elucidate the roles and interactions between microorganisms in the environment of interest. Metagenomic data analysis is challenging for beginners without appropriate guidance and resources. This discipline mandates not only robust knowledge of bioinformatics but also proficiency in programming

languages, necessitating extensive training and practice. ChatGPT-4, developed by OpenAI, is a language model trained on a diverse array of data. It can generate programming code based on structured natural language instructions. Thus, ChatGPT-4 is a potential facilitator for metagenomic study, beneficial for both beginners and professionals. It aids in writing and optimizing programming code, thereby accelerating and enhancing the data analysis workflow. In this case study, we demonstrated a prompt pipeline employing structed natural language to instruct ChatGPT-4 in generating code for metagenomic data analysis. The pipeline consists of eight tasks. We found ChatGPT-4, when provided with clear, structured instructions, can produce functional programming code to complete varies steps involved in a metagenomic analysis. Additionally, ChatGPT-4 can self-correct its code when presented with error messages or further natural language instructions. The robustness of the pipeline was validated through internal and external testing. Through this case study, we demonstrated ChatGPT-4 as a potential assistant for shotgun metagenomic data analysis. We illustrated that, given structured natural language, ChatGPT-4 can autonomously generate programming code that delivers correct results.

Keywords: ChatGPT-4, Prompt Engineering, Metagenomics, Microbiome

---

Abstract ID: 1394
Cancer Comprehend Annotation – a pipeline for cancer phenotype and clinical extraction

Thanh Duong[1], Phillip Szepietowski[2], Thanh Thieu[1]

[1]Department of Machine Learning, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States
[2]Department of Health Data Services, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States.

Abstract
Information extraction from clinical text is needed to comprehend patient conditions and determine anticancer treatment. Existing NLP systems such as Amazon Comprehend Medical, CLAMP, and DeepPhe are proprietary and suboptimal due to the shift in textual distribution and expression of cancer phenotypes. We introduce OncoNLP, a natural language processing toolkit that comprises deep neural network Bidirectional Encoder Representations from Transformers (BERT) models designed to extract cancer phenotype and related biomedical information. The toolkit contains two primary components: a Biomedical BERT (BiomedBERT) model to extract general medical information, and a Cancer BERT (CaBERT) model to identify primary tumor site and histology. BiomedBERT is a named entity recognition (NER) model that performs sequence labeling on three medical events: Problem, Test and Treatment. CaBERT is a question-answering model that was trained and tested on pathological notes at Moffitt. We evaluate performance of BiomedBERT on Informatics for Integrating Biology & the Bedside (i2b2) dataset against Amazon Comprehend Medical and CLAMP. BiomedBERT outperforms both other methods with exact matching F1-score at 88.5%, while CLAMP attains 88.1% and Amazon Comprehend attains 85.5%. Next, we evaluate CaBert on a Moffitt held-out test dataset with over 2000 clinical notes against DeepPhe. CaBert outperforms Deepphe on exact matching of tumor site at 73.2% F1-score versus 28.1%, and exact matching of tumor histology at 85.3% F1-score versus 22.4%. Lastly, we use John Snow Lab to present extracted information medical note that could be reviewed and modified. In future work, we plan to expand

capabilities of OncoNLP to detect recurrence and progression, as well as testing its applicability in clinical settings.

Keywords: Natural language processing, cancer phenotypes, oncology informatics.

Abstract ID: 1986
Genomic disparities between cancers in adolescent and young adults and in older adults

Xiaojing Wang[1,2], Anne-Marie Langevin[3], Peter Houghton[1,4], Siyuan Zheng[1,2]

[1]Greehey Children's Cancer Research Institute, UT Health San Antonio, TX, USA;
[2]Department of Population Health Sciences, UT Health San Antonio, TX, USA;
[3]Department of Pediatrics, UT Health San Antonio, TX, USA;
[4]Department of Molecular Medicine, UT Health San Antonio, TX, USA.

Abstract
Cancers cause significant mortality and morbidity in adolescents and young adults (AYAs), but their biological underpinnings are incompletely understood. In this study, we analyze clinical and genomic disparities between AYAs and older adults (OAs) in more than 100,000 cancer patients. We find significant differences in clinical presentation between AYAs and OAs, including sex, metastasis rates, race and ethnicity, and cancer histology. To identify genetic differences, we built logistic regression models controlling for clinical and genomic confounders. We find in most cancer types, AYA tumors show lower mutation burden and less genome instability. Accordingly, most cancer genes show less mutations and copy number changes in AYAs, including the noncoding TERT promoter mutations. However, CTNNB1 and BRAF mutations are consistently overrepresented in AYAs across multiple cancer types. AYA tumors also exhibit more driver gene fusions that are frequently observed in pediatric cancers. However, AYA tumors are not a simple extension of pediatric cancer. For instance, in brain tumor, common genetic events do not necessarily show changes correlated with patient age groups. We find that histology is an important contributor to genetic disparities between AYAs and OAs. Mutational signature analysis of hypermutators shows stronger endogenous mutational processes such as MMR-deficiency but weaker exogenous processes such as tobacco exposure in AYAs. Finally, we demonstrate a panoramic view of clinically actionable genetic events in AYA tumors. In conclusion, this systematical analysis reveals genetic and clinical disparities between OAs and AYAs with cancer.

Keywords: Adolescents and Young Adults, Cancer Disparity, Panel Sequencing, Age

---

Abstract ID: 1249
TSSr: an R package for comprehensive analyses of TSS sequencing data

Zhaolian Lu[1], Keenan Berry[2], Zhenbin Hu[1], Yu Zhan[1], Tae-Hyuk Ahn[2,3], Zhenguo Lin[1,2]

[1]Department of Biology, Saint Louis University, St. Louis, MO 63103, USA;

[2]Program of Bioinformatics and Computational Biology, Saint Louis University, St. Louis, MO 63103, USA; [3]Department of Computer Sciences, Saint Louis University, St. Louis, MO 63103, USA

Abstract
Transcription initiation is regulated in a highly organized fashion to ensure proper cellular functions. Accurate identification of transcription start sites (TSSs) and quantitative characterization of transcription initiation activities are fundamental steps for studies of regulated transcriptions and core promoter structures. Several high-throughput techniques have been developed to sequence the very 5' end of RNA transcripts (TSS sequencing) on the genome scale. Bioinformatics tools are essential for processing, analysis, and visualization of TSS sequencing data. Here, we present TSSr, an R package that provides rich functions for mapping TSS and characterizations of structures and activities of core promoters based on all types of TSS sequencing data. Specifically, TSSr implements several newly developed algorithms for accurately identifying TSSs from mapped sequencing reads and inference of core promoters, which are a prerequisite for subsequent functional analyses of TSS data. Furthermore, TSSr also enables users to export various types of TSS data that can be visualized by genome browser for inspection of promoter activities in association with other genomic features, and to generate publication-ready TSS graphs. These user-friendly features could greatly facilitate studies of transcription initiation based on TSS sequencing data. The source code and detailed documentations of TSSr can be freely accessed at https://github.com/Linlab-slu/TSSr.

Keywords: TSS, promoter, R package, gene expression, CAGE

Abstract ID: 1579
Building the Human Ensemble Cell Atlas and Learning the Underlying Unified Coordinate System

Xuegong Zhang[1,2]

[1]MOE Key Lab of Bioinformatics and Bioinformatics Division of BNRIST, Department of Automation, Tsinghua University, Beijing, China;
[2]School of Life Sciences and School of Medicine, Tsinghua University, Beijing, China.

Abstract
Building an atlas of all human cell types with their gene expression properties at single-cell resolution can provide a fundamental reference to future human biology and medicine. We built the first human ensemble cell atlas hECA using cell-centric assembly from scattered data sources, and invented the new paradigm of "in data" cell experiment by taking the cell atlas as a virtual human body composed of in data cells. Cells exhibit multifaceted heterogeneity at multiple scales. Finding the major attributes that can be used to index or sort the cells with regard to different aspects of the heterogeneity is important for building and utilizing cell atlases. For this purpose, we developed a multidimensional coordinate system UniCoord for different physical and biological attributes of cells by adopting a supervised variational autoencoder (VAE) neural network model. We trained UniCoord on the first cell-centric assembled human single-cell atlas hECA to make it represent the diversity of healthy human cells. Experiments showed that UniCoord is able to capture key cellular features of spatial, temporal and functional gradients from massive data. These features are sufficient for accurate data reconstruction and label identification and can be interpolated to predict intermediate cell states between two discrete cell groups for studying cell state transition and cell type

differentiation. UniCoord provides a prototype for a learnable universal coordinate framework for organizing sophisticated cell atlases to enable better analyzing the highly orchestrated functions and multifaceted heterogeneities of diverse cells of an organ, a system or the whole human body.

Keywords: Cell atlas; Single-cell RNA-seq data; Machine learning; *in data* cell experiment.

Abstract ID: 1617

Deep Transfer Learning of Cancer Drug Responses by Integrating Bulk and Single-cell RNA-seq data

Junyi Chen[1,*], Xiaoying Wang[2,*], Anjun Ma[1,3,$], Qi-En Wang[4], Bingqiang Liu[2], Lang Li[1], Dong Xu[5], Qin Ma[1,3,$]

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA
[2] Department of Mathematics, Shandong University, Shandong 250100, China
[3] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA.
[4] Department of Radiation Oncology, Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
[5] Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[*] These authors contributed equally
[$] To whom correspondence should be addressed

Abstract

Drug screening data from massive bulk gene expression databases can be analyzed to determine the optimal clinical application of cancer drugs. The growing amount of single-cell RNA sequencing (scRNA-seq) data also provides insights into improving therapeutic effectiveness by helping to study the heterogeneity of drug responses for cancer cell subpopulations. Developing computational approaches to predict and interpret cancer drug response in single-cell data collected from clinical samples can be very useful. We propose scDEAL, a deep transfer learning framework for cancer drug response prediction at the single-cell level by integrating large-scale bulk cell-line data. The highlight in scDEAL involves harmonizing drug-related bulk RNA-seq data with scRNA-seq data and transferring the model trained on bulk RNA-seq data to predict drug responses in scRNA-seq. Another feature of scDEAL is the integrated gradient feature interpretation to infer the signature genes of drug resistance mechanisms. We benchmark scDEAL on six scRNA-seq datasets and demonstrate its model interpretability via three case studies focusing on drug response label prediction, gene signature identification, and pseudotime analysis. We believe that scDEAL could help study cell reprogramming, drug selection, and repurposing for improving therapeutic efficacy.

Keywords: Single-cell RNA-seq, cancer drug response, deep transfer learning

---

Abstract ID: 1816

Decentralization of Brain age Estimation with Structural Magnetic Resonance Imaging Data

Sunitha Basodi[1], Rajikha Raja[2], Bhaskar Ray[1,3], Harshvardhan Gazula[4], Jingyu Liu[1,3], Eric Verner[1] and Vince D. Calhoun[1,3,5]

[1] Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA
[2] St. Jude Children's Research Hospital, Memphis, TN, USA
[3] Department of Computer Science, Georgia State University, Atlanta, GA, USA
[4] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[5] Department of Psychology, Georgia State University, Atlanta, GA, USA

Abstract:
Brain age estimation is a widely used approach to evaluate the impact of various neurological or psychiatric brain disorders on the brain developmental or aging process [Jónsson et al., 2019, Reeve et al., 2014]. Current studies show that neuroimaging data can be used to predict brain age, as it captures structural and functional changes that the brain undergoes during development and the aging process [Cole et al., 2018, Liem et al., 2017]. Although access to large amounts of neuroimaging data helps build better models and validate their effectiveness, researchers often have limited access to brain data because of its challenging and expensive acquisition process. This data is not always sharable due to privacy restrictions. Decentralization provides a way which does not require data exchange between the multiple involved groups [Aledhari et al., 2020, Li et al., 2020]. In this work, we propose a decentralized approach for brain age prediction in which a decentralized model is built by using the model parameters of the participating sites without sharing its data. In the first training step, all the participating sites train a support vector regression (SVR) model locally with their data and transfer the weight vectors of the locally learned models to one site (allocated as the master site during setup). This master site averages all these weight vectors and uses the averaged weight vector to transform its data into a new feature space. This modified data is then used to train a decentralized SVR model at the master site (see Fig. 1) and the learnt parameters of the decentralized SVR model are sent to all other sites. We use COINSTAC platform [Plis et al., 2016, White et al., 2020] to implement our decentralized SVR models. We evaluate our models using the features extracted from structural Magnetic Resonance Imaging (MRI) data. We compare performance of the models trained with three different data sampling strategies and showed that decentralized models have similar performance to their corresponding centralized (trained with all the data in one location) models. The key benefit of our approach is that it encourages collaboration by allowing different research groups to readily participate in larger brain age analysis without worrying about their data-sharing policies or data transmission. This work has been published in Neuroinformatics journal [Basodi et al., 2022].

Keywords: Brain Age Prediction, Decentralized, Federated, COINSTAC, Neuroimaging

---

Abstract ID: 1278
An integrative study to identify the link between dysregulated intercellular signaling and genetic variants in Alzheimer's disease

Andi Liu[1,2], Xiaoyang Li[2,3], Brisa S Fernandes[2], Yulin Dai[2], Zhongming Zhao[1,2,4,*]

[1]Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA;
[2]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA;
[3]Biostatistics & Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA;
[4]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract
Alzheimer's disease (AD) is a complex and debilitating neurodegenerative disease affecting over 40 million people worldwide. The onset and progression of AD are heavily influenced by genetic variants, which have been the focus of extensive research in recent years. Researchers have identified over 75 genetic variants linked to AD through genome-wide association studies (GWASs) and have used single-nuclei RNA sequencing (snRNA-seq) data to identify changes in gene expression and dysregulated intercellular signals. However, whether and how AD-associated genetic variants manifest their impacts on intercellular signaling and functional pathways in disease progression remains poorly understood. In this study, we aim to uncover and validate high confident dysregulated intercellular signals associated with AD by integrating snRNAseq, GWAS, and whole genome sequencing (WGS) data. Specifically, we conducted an in-depth analysis on publicly available transcriptome profiles of 69,787 single nuclei, which are obtained from prefrontal cortex (PFC) postmortem samples of 24 AD cases and matched controls from the Religious Orders Study and Memory and Aging Project (ROSMAP). Using CellChat, a computational tool for inferring intercellular communication networks based on prior knowledge of ligand-receptor (LR) pairs and signaling pathways, we identified 315 unique dysregulated LR pairs in AD. Among them, 247 incoming/outgoing signals were decreased, and 68 signals were increased across various cell types. To corroborate these findings, we performed parallel pathway-level analyses on 842 Gene Ontology (GO) functional pathways containing predicted dysregulated LR pairs. We employed MAGMA, a state of art pathway analysis tool leveraging population-level (GWAS statistics) genetic information, and PRSet, a novel pathway-level polygenic risk score (PRS) tool leveraging individual-level (WGS) genetic information, to prioritize AD-associated GO pathways. As a result, we validated 110 LR pairs involved in the identified GO pathways. We highlighted seven LR pairs that were validated in both analyses, such as APOE-SORL1, APOE-LRP1 and LRPAP1-LRP1, which are well known signaling axes associated with amyloid precursor protein regulation and process, among others. Furthermore, we observed cell-type specificity of identified LR pairs, particularly in astrocytes, microglia, and excitatory neurons. To conclude, the enriched genetic risks linked to the cellular cross-talks among astrocytes, microglia, and excitatory neurons were identified, providing new insights into the potential therapeutic targets involved in dysregulated cell-cell communication in AD.

Keywords Alzheimer’s disease; single nuclei RNA-seq; cell-cell communication; GWAS; whole genome sequencing

---

Abstract ID: 1915

A massive proteogenomic screen identifies thousands of novel peptides from the human "dark" proteome

Xiaolong Cao[1,2], Siqi Sun[2], Jinchuan Xing[2,*]

[1] Division of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong 510280, China;
[2] Department of Genetics, Human Genetic Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

Abstract
Although the human gene annotation has been continuously improved over the past two decades, numerous studies demonstrated the existence of a "dark proteome", consisting of proteins that were critical for biological processes but not included in widely-used gene models. The Genotype-Tissue Expression (GTEx) project generated more than 15,000 RNA sequencing datasets from multiple tissues, which modeled 30 million transcripts in the human genome. To provide a resource of high-confidence novel proteins from the dark proteome, we screened 50,000 mass spectrometry runs to identify proteins translated from the GTEx transcript model with proteomic support. We also integrated 3.8 million common genetic variants from the gnomAD database to improve peptide identification. With a stringent standard, we identified more than 24,000 novel peptides originating from the dark proteome. Our method showed remarkable potential in identifying novel proteins from the dark proteome. The findings will improve our understanding of coding genes and facilitate genomic data interpretation in biomedical research.

Keywords: Proteogenomics, noncanonical open reading frames, mass spectrometry, RNA-seq

---

Abstract ID: 1895
Mutated processes predict immune checkpoint inhibitor therapy benefit in metastatic melanoma

Andrew Patterson[1], Noam Auslander[2]

[1]Genomics and Computational Biology Graduate Group, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[2]Program in Molecular and Cellular Oncogenesis, The Wistar Institute, Philadelphia, PA, 19104, USA

Abstract
Immune Checkpoint Inhibitor (ICI) therapy has revolutionized treatment for advanced melanoma; however, only a subset of patients benefit from this treatment. Despite considerable efforts, the Tumor Mutation Burden (TMB) is the only FDA-approved biomarker in melanoma. Yet, the mechanisms underlying TMB association with prolonged ICI survival are not entirely understood and may depend on numerous confounding factors. To identify more interpretable ICI response biomarkers based on tumor mutations, we train classifiers using mutations within distinct biological processes. We evaluate a variety of feature selection and classification methods and identify key mutated biological processes that provide improved predictive capability compared to the TMB. Feature selection methods evaluated are greedy and randomized forward selection algorithms and genetic algorithms. Classification methods considered are

Random Forest and Gradient Boosting tree methods, Feedforward, and Long Short-Term Memory Neural Networks. We found that Random Forest has the most generalizable performance for ICI prediction through mutated processes. The best performing mutated processes are leukocyte and T-cell proliferation regulation, which demonstrate stable predictive performance across different data cohorts of melanoma patients treated with ICI. We show that this approach uncovers specific genes within a process associated with ICI response or resistance. Finally, we show that our model predicts survival in both ICI treated and untreated patients. This study provides biologically interpretable genomic predictors of ICI response with substantially improved predictive performance over the TMB.

Keywords: Machine Learning, Melanoma, Immune Checkpoint Inhibitors

---

Abstract ID: 1662
HiC4D: Forecasting spatiotemporal Hi-C data with residual ConvLSTM

Tong Liu[1], Zheng Wang[1]

[1]Department of Computer Science, University of Miami, Coral Gables, Florida, USA

Abstract
The Hi-C experiments have been extensively used for the studies of genomic structures. In the last few years, spatiotemporal Hi-C has largely contributed to the investigation of genome dynamic reorganization. However, computationally modeling and forecasting spatiotemporal Hi-C data still have not been seen in the literature. We present HiC4D for dealing with the problem of forecasting spatiotemporal Hi-C data. We designed and benchmarked a novel network and named it residual ConvLSTM (ResConvLSTM), which is a combination of residual network and convolutional long short-term memory (ConvLSTM). We evaluated our new residual ConvLSTM networks and compared them with another five methods including a naive network (NaiveNet) that we designed as a baseline method and four outstanding video-prediction methods from the literature: ConvLSTM, spatiotemporal LSTM (ST-LSTM), self-attention LSTM (SA-LSTM), and simple video prediction (SimVP). We used eight different spatiotemporal Hi-C datasets for the blind test, including two from mouse embryogenesis, one from somatic cell nuclear transfer (SCNT) embryos, three embryogenesis datasets from different species, and two non-embryogenesis datasets. Our evaluation results indicate that our residual ConvLSTM networks almost always outperform the other methods on the eight blind-test datasets in terms of accurately predicting the Hi-C contact matrices at future time steps. Our benchmarks also indicate that all of the nine methods can successfully recover the boundaries of topologically associating domains (TADs) called on the experimental Hi-C contact matrices. Taken together, our benchmarks suggest that HiC4D is an effective tool for predicting spatiotemporal Hi-C data. HiC4D is publicly available at http://dna.cs.miami.edu/HiC4D/.

Keywords: 4D genome, spatiotemporal Hi-C, deep learning, convolutional long short-term memory

---

Abstract ID: 1803
Integrating Hydrogen Bonding Information into Graph Neural Networks for Protein Structure Classification

Yi-Shan Lan[1], Tsung-Yi Ho[2]

[1]Department of Computer Science, Institute of Information Systems & Applications, National Tsing Hua University, Hsinchu 30013, Taiwan
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong

Abstract
3D protein structures are commonly analyzed using a distance map, which is widely employed in protein prediction tasks. However, this distance matrix lacks crucial chemical features. To enhance our understanding of the chemical aspects within proteins, our aim is to enrich the datasets with these properties. We propose the graph-based chemical bond dataset, which intuitively showcases the chemical bonds within the molecule. Our methodology leverages Graph Neural Networks (GNN) to effectively integrate chemical properties and bonding information. Specifically, we incorporate hydrogen bond network by protonating the protein coordinates. In our experiments, our approach achieved a remarkable accuracy of 94% in protein structure-based classification, encompassing 1161 classes of the Structural Classification of Proteins-Extended (SCOPe). This surpasses the performance of other deep learning methods using structures and establishes a new state-of-the-art. Hydrogen bonds in proteins play a pivotal role in specific molecular recognition. The inclusion of hydrogen bonding data significantly enhances the representation of protein structure and function. Through the utilization of the GNN model, we gain valuable insights into the intricate relationship between protein structure and hydrogen bonding.

Keywords: Graph-based approach, Hydrogen bonding networks, Graph neural networks, Structural Classification of Proteins-Extended (SCOPe)

Poster Session

Monday, July 17, 2023
5:40-6:40 PM
Room: St. Petersburg II, III

Abstract ID: 1002
Identification of Key Biomarkers Associated with Ductal Breast Cancer in Spatial Transcriptomics Data

Ellie Xi[1], Tutu Hu[2], Chloe Yu[3], Lulu Shang[4], Xiang Zhou[4], Allen Bai[5,6]

[1]BASIS Independent Silicon Valley, 1290 Parkmoor Ave, San Jose, CA 95126 USA;
[2]Tabor Academy, 66 Spring Street, Marion, MA 02738 USA;
[3]Union County Academy for Allied Health Sciences, 1776 Raritan Road, Scotch Plains, NJ 07076 USA;
[4]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 USA;
[5]Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197 USA;
[6]Next-Gen Intelligent Science Training, Ann Arbor, MI 48105 USA.

Abstract: Refer to Session "Artificial Intelligence on Big Data: Promise for Early-stage Trainees".

Abstract ID: 1003
Shared genetic basis informs the roles of polyunsaturated fatty acids in brain disorders

Huifang Xu[1], Yitang Sun[1], Michael Francis[2], Claire Cheng[1], Nitya Modulla[1], Kaixiong Ye[1,2]

[1]Department of Genetics, University of Georgia, Athens, Georgia, USA;
[2]Institute of Bioinformatics, University of Georgia, Athens, Georgia, USA;

Abstract: Refer to Flash Talk Session

Abstract ID: 1010
Machine Learning-based staging of kidney cancer using microRNA expression profiles

Ming-Ju Tsai[1,2], Nikhila Aimalla[3], Sanjay K Shukla[4], Rohit Sharma[5] and Srinivasulu Yerukala Sathipati[4*]

[1]Hinda and Arthur Marcus Institute for Aging Research at Hebrew Senior Life, Boston, MA, USA
[2]Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA
[3]Department of Internal Medicine-Pediatrics, Marshfield Clinic Health System, Marshfield, WI [4]Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, USA
[5]Department of Surgical Oncology, Marshfield Clinic Health System, Marshfield, WI, USA
[*]Corresponding author

Abstract
Introduction: Kidney cancer presents a significant global health challenge due to its high prevalence. The objective of this study was to develop machine learning (ML) models to distinguish early and advanced stages of kidney clear cell renal cell carcinoma (KCRC) and kidney papillary renal cell carcinoma (KPRC). **Materials and Methods:** MicroRNA (miRNA) expression profiles of 492 patients with KCRC and KPRC were extracted from The Cancer Genomic Atlas database. An evolutionary modeling algorithm was employed to conduct feature selection with state-of-the-art ML methods, including Catboost, XGBoost, LightGBM, Gradient Boosting, Logistic Regression, Ridge Classifier, Support Vector Machine, Linear Discriminant Analysis (LDA), Random Forest, and Extra Trees. **Results**: Based on 10-fold cross-validation (CV) results, the best KCRC model was obtained using 23 miRNAs with the LDA classifier, achieving an accuracy of 77%, an Area Under the Receiver Operating Characteristic (AUROC) curve of 0.82, and an Area Under the Precision-Recall Curve (AUPRC) of 0.75. This model was also validated on an independent test dataset, yielding an accuracy of 82%, an AUROC of 0.88, and an AUPRC of 0.82 in distinguishing early and advanced stages of KCRC. Similarly, for KPRC, the LDA classifier with 21 miRNAs demonstrated strong performance, with an accuracy of 83%, an AUROC of 0.88, and an AUPRC of 0.78 in the 10-fold CV. The independent test dataset further confirmed its efficacy, showing an accuracy of 92%, an AUROC of 0.89, and an AUPRC of 0.87, in distinguishing early and advanced stages of KPRC. Notably, two miRNAs, hsa-miR-101-3p and hsa-miR-301a-3p, were identified as common biomarkers in both types

of kidney cancer. Additionally, we investigated the relationship between miRNA signatures and target genes by examining their roles in renal-specific cell lines, as well as analyzing Gene Effect Scores obtained from CRISPR-based knockout screenings. **Conclusion**: In conclusion, this study successfully developed ML models that accurately stage KCRC and KPRC based on miRNA expression profiles. The comprehensive pathway enrichment analyses and identification of cell line-specific target genes shed light on the underlying mechanisms of KCRC and KPRC. These findings highlight the potential of ML methods in improving cancer staging processes, providing clinicians with valuable tools for more precise staging and prognostic treatment decisions. In the future, our focus will be on developing ML-based prognostic methods specifically designed for kidney cancers.

Keywords: Clear cell renal cell carcinoma (KCRC), Papillary renal cell carcinoma (KPRC), Machine-learning methods, miRNA signature

---

Abstract ID: 1012
A Super-Enhancer-Based Fingerprint of Human Cancers

Xiang Liu[1], Nancy Gillis[2], Chang Jiang[3], Anthony McCofie[1], Timothy I Shaw[1], Aik-Choon Tan[4], Bo Zhao[5], Lixin Wan[6], Derek R Duckett[7], and Mingxiang Teng[1]*

[1]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA.
[2]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL 33612, USA.
[3]Department of Metabolism and Physiology, Moffitt Cancer Center, Tampa, FL 33612, USA.
[4]Department of Oncological Sciences, Huntsman Cancer Institute, The University of Utah, Salt Lake City, UT 84112, USA.
[5]Division of Infectious Disease, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[6]Department of Molecular Oncology, Moffitt Cancer Center, Tampa, FL 33612, USA.
[7]Department of Drug Discovery, Moffitt Cancer Center, Tampa FL 33612, USA.

Abstract
Super enhancers (SEs) have been found as critical regulators in human cancers. We have previously reported that, instead of treating each SE as a single unit, considering the dynamics of their internal components (i.e., constituent enhancers or CEs), provides better characterization of SE alterations. Here, we extend the analysis to pan-cancer datasets to identify cell-specific SEs in human cancers, by fully utilizing SE internal dynamics and computationally modeling active enhancers. H3K27Ac ChIP-seq data of NCI-60 cells were utilized to measure enhancer activity in 28 cancers. SE candidates presented in at least one cell line were first identified using ROSE. Promoter regions were excluded. Due to the varied overall activities of enhancers, the measurements of an active enhancer with H3K27Ac ChIP-seq data differ across enhancers. In other words, a low ChIP-seq measurement might not necessarily correspond to inactive enhancer status. For this reason, we developed a mixture model to better separate active cancer cell types from inactive cells for each enhancer, by fully stratifying ChIP-seq signals across NCI-60 cell lines. We benchmarked the superiority of the refined enhancer activity by comparing it to the commonly defined enhancer status by ChIP-seq peak calling, based on public 3D chromatin looping and massive parallel reporter assay data. Based on the refined CE activity, we further identified cell/cancer-specific CEs/SEs as

an epigenomic fingerprint for human cancers. We found that the refined CE statuses better accommodate with gene regulation (by ChIA-PET looping) and experimental enhancer activity (by massive parallel reporter assay). The refined CE statuses provide better characterization of cancer identities, compared to those by genome-wide enhancers or non-SE enhancers. This suggests the feasibility a fingerprint of cancers using these CEs/SEs. We built the fingerprint using cancer-specific CEs/SEs that were prioritized by their capability in charactering cancer identities. We implemented the fingerprint with an R package (cSEAdb) to facilitate query and exploring.

Keywords: Pan-cancer, Super-enhancer, Cancer epigenomic

Abstract ID: 1025
Comparisons of Coronavirus Spike Proteins and the Mutation Effects on Virus-Host Interaction

Crystal Teng[1], Vidhyanand Mahase[2], Adebiyi Sobitan[2], Shaolei Teng[2]

[1]Northwest High School, Germantown, MD, 20874 [2]Department of Biology, Howard University, Washington, D.C., 20059 USA

Abstract: Refer to Session "Artificial Intelligence on Big Data: Promise for Early-stage Trainees".

Abstract ID: 1043
Glucose-lowering and hypolipidimic activities of ethanolic extract from Aloe vera in STZ-induced diabetic rats

P. Srikrishna[1], Sudha, Dr .Prem,A. Rojarani*

[1, *]Department of Genetics , Osmania University, Hyderabad.

ABSTRACT
Aloe vera is a traditionally, and a medicinal plant has been employed to treat skin problems (burns, wounds, and anti-inflammatory processes). Moreover, Aloe vera has shown other therapeutic properties including anticancer, antioxidant, ant diabetic, and anti hyper lipidemic. which has tremendous biological activities In this study aimed to asses the glucose –lowering and hyper lipidimic effects of ethanolic extract from Aloe vera. (AV) in Streptozotocin (STZ) – Induced mice. Aloe vera leaves were extracted in hot water and ethanol was mixed .I used most convenient method Shodex sugar KS 804 Column chromatography. The supernatant was collected and mixed with absolute ethanol at a ratio of1:3 for 12 h. The crude extracts are obtained by ethanol precipitation method .To asses the hyperlipidimic activity in induced diabetic mice used crude extract given to mice through oral in some cases with STZ .and other drugs through subcutaneous way. For 30 days with different concentration of drugs in form of slaine or liquid by using micro syringe. After completion of treatment with aloevera extract diabetic induced mice showed normal glucose level.

Abstract ID: 1056

f-divergence based generative adversarial imputation method for enhanced single-cell RNAseq data analysis

Tong Si[1], Zackary Hopkins[2], John Yanev[2], Jie Hou[2] , Haijun Gong[1]
[1]Department of Mathematics and Statistics, Saint Louis University, Saint Louis, MO, USA; [2]Department of Computer Sciences, Saint Louis University, Saint Louis, MO, USA

Abstract
Single-cell RNA sequencing (scRNA-seq) technology allows researchers to analyze thousands of individual cells simultaneously, providing insights into cellular heterogeneity. Analyzing single-cell RNA sequencing data enhances our understanding of cellular diversity and aids in developing personalized therapies. Missing values, known as dropouts, are common in single-cell RNA sequencing data due to technical limitations. The abundance of missing values presents a significant challenge for downstream analysis, making the analysis of scRNA-seq difficult.
Traditional methods employed for imputing missing values in other types of data struggle when confronted with the high-dimensional nature of scRNA-seq datasets and the complex statistical properties of these missing values. Moreover, their performance in imputing missing values diminishes as the missing rates increase, rendering them less effective for handling datasets with substantial dropout proportions. To address the challenges of traditional methods in scRNA-seq data imputation, we propose a novel f-divergence-based generative adversarial imputation network, called sc-fGAIN. Unlike traditional methods, sc-fGAIN utilizes a deep generative model that trains two neural networks to generate imputed values without relying on any assumptions. Our theoretical analysis identified four f-divergence functions, namely cross-entropy, forward Kullback-Leibler, reverse KL, and Jensen-Shannon, that can be effectively used as the adversarial loss function in the sc-fGAIN architecture. Crucially, our mathematical proofs have confirmed that the distribution of imputed data using the sc-fGAIN method aligns with the distribution of the original data. This ensures that the imputed values retain the intrinsic characteristics of the underlying cellular diversity captured by the scRNA-seq measurements. Real scRNA-seq data analysis has shown that, compared to most traditional methods, the imputed values generated by sc-fGAIN have a smaller root-mean-square error, and it is robust to varying missing rates, moreover, our sc-fGAIN can reduce imputation bias by producing smaller standard deviations of gene expression values. In conclusion, according to our theoretical analysis, it has been proved the generative adversarial imputation network method remains valid when the adversarial loss function is defined using four different f-divergence functions, furthermore, experiments based on the real data analysis also demonstrated that the sc-fGAIN method could give competitive results compared with traditional imputation methods. The flexibility offered by the f-divergence formulation allows sc-fGAIN to accommodate various types of data, making it a more universal approach for imputing missing values of scRNA-seq data. All in all, our findings highlight the promising potential of sc-fGAIN in enhancing the quality of single-cell RNA sequencing data and facilitating more precise and reliable downstream analyses.

Keywords: Single cell, missing value imputation, f-divergence, generative adversarial networks

---

Abstract ID: 1059
Pan-cancer analysis of metabolic shifts via flux estimation analysis

Kevin Hu[1,2], Alex Lu[1,3], Grace Yang[1,2], Shaoyang Huang[1,2], Pengtao Dang[1,4], Yijie Wang[1,5], Haiqi Zhu[1,5], Sha Cao[1,6], Chi Zhang[1,7]

[1]Center for Computational Biology and Bioinformatics, [6]Department of Biostatistics, [7]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA.
[2]Carmel High School, Carmel, IN, USA. [3]Park Tudor School, Indianapolis, IN, USA.
[4]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, USA.
[5]Department of Computer Science, Indiana University, Bloomington, IN, USA

Abstract: Refer to Session "Artificial Intelligence on Big Data: Promise for Early-stage Trainees".

Abstract ID: 1078
Identifying Chronic Tic Disorder subtypes using clinical diagnostic data

Subramanian, Krishnamurthy[1], Tourette International Collaborative Genetics (TIC Genetics) group, and Jinchuan Xing[1]

[1] Department of Genetics and the Human Genetics Institute of New Jersey, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, USA

Abstract
Chronic Tic Disorder (CTD), including Tourette's syndrome (TS) and other tic disorders, is a heterogeneous, childhood-onset neurodevelopmental disorder. CTD is characterized by the presence of motor and/or vocal tics and it affects 1-3% of the population. About 88% of the patients have other neurodevelopmental disorder comorbidities, suggesting shared genetic risk factors for these disorders. Because of high level of heterogeneity and comorbidities, we hypothesize that distinct subtypes exist among CTD patients. Here we identified CTD subtypes and evaluated their discriminatory factors among patients in the Tourette International Collaborative Genetics (TIC Genetics) study. Using Hierarchical Ascendant Clustering, Finite Mixture Modeling, and Bayesian Hierarchical Clustering, we analyzed the TIC Genetics diagnostic data (18 variables) for 1900 CTD patients. These methods identified six distinct clusters: 1. All subjects with CTD but did not meet the diagnosis criteria of TS. These subjects also have a lower prevalence of hispanics and lower prevalence of comorbidities; 2. Subjects with TS. These subjects also have a lower prevalence of hispanics and lower prevalence of comorbidities; 3. Hispanic subjects; 4. Subjects with CTD combined subtype from multi-birth, likely due to environmental conditions during/after birth; 5. Subjects with TS and Trichotillomania and high prevalence of attention deficit obsessive compulsive disorder (OCD); and 6. Subjects with TS with higher prevalence of OCD and ADHD. In summary, our results show that distinct clusters can be identified among CTD patients based on clinical data. In the future, we will conduct stratified analysis of genetic data (e.g., microarray and whole exome sequencing) based on these subtypes to determine the genetic etiology of the subtypes.

Keywords: machine learning, Chronic Tic Disorder, Tourette's syndrome, Neurodevelopmental disorder, Clinical subtypes

---

Abstract ID: 1117
Dynamic functional network connectivity separates patients from control and their relatives

Mohammad SE. Sendi[1,2], Hossein Dini[3], Vince D. Calhoun[2,4]

[1] McLean Hospital and Harvard Medical School, Boston, MA, USA;
[2] Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA;
[3] Aalborg University, Copenhagen. Denmark;
[4] Department of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA, USA

Abstract
Background: Dynamic functional network connectivity (dFNC) has shown promise in distinguishing between patient groups and healthy controls (HC). However, the challenge lies in identifying the distinct patterns of mental diseases, including schizophrenia (SZ), bipolar disorder (BP), and schizoaffective disorder (SAD). Currently, there is limited understanding regarding the overlap between these conditions. Therefore, this study aims to explore whether dFNC features can effectively differentiate patients with psychotic disorders and their relatives from HC. By investigating the unique neural connectivity patterns associated with these disorders, we can potentially enhance our understanding of their underlying mechanisms and improve diagnostic accuracy. **Methods:** We used resting-state fMRI, demographics, and clinical information from the Bipolar–Schizophrenia Network on Intermediate Phenotypes cohort. The data includes three groups of patients with schizophrenia (SZP, N=102), bipolar (BPP, N=102), and schizoaffective (SADP, N=102), their relatives SZR (N=102), BPR (N=102), SADR (N=102), and HC (N=118) groups. No significant age and sex differences were observed across groups. After estimating each individual's dFNC, we put them into three identical states using the k-means clustering approach. Next, we estimated each state's occupancy rate (OCR) for each participant. Finally, we compared OCR statistically across patients, their relatives, and HC groups. **Results:** We found OCR differentiates patients from HC in SZ (corrected $p<0.001$), BP (corrected $p=0.045$), and SAD (corrected $p= 0.001$). While it only separated relatives from patients in SZ (corrected $p=0.021$). Additionally, we found dFNC feature separated BPP from SZP (corrected $p= 0.012$) and SADP (corrected $p= 0.017$) significantly. However, it could not separate SZR, BPR, and SADR from each other. **Conclusion:** Our findings provide compelling evidence that the dFNC feature can effectively differentiate patients from both the control group and their relatives. Building upon these results, our future research will delve deeper into exploring additional dFNC features with the aim of further distinguishing patients with psychotic disorders from the control group and their relatives. By expanding our analysis, we hope to gain a more comprehensive understanding of the underlying neural mechanisms involved in these conditions, ultimately contributing to improved diagnostic capabilities.

Keywords: dynamic functional network connectivity, neuropsychiatric disorders, resting-state functional magnetic resonance imaging

---

Abstract ID: 1118
An active learning framework for personalized deep brain stimulation

Mohammad SE. Sendi[1,2], Robert E.Gross[3], Vince D. Calhoun[2,4]

[1] McLean Hospital and Harvard Medical School, Boston, MA, USA;
[2] Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA;
[3] Department of Neurosurgery at Emory University, Atlanta, GA, USA;
[4] Department of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Background: To personalize deep brain stimulation (DBS), it is crucial to establish a connection between DBS parameters and an individual's neural response. The current approach, which relies on random sampling (RS), is not only time-consuming and costly but also impractical in a clinical setting. Consequently, it hinders the exploration of novel settings when faced with challenging situations. To tackle this issue, we have developed a novel algorithmic framework centered around active learning (AL). This framework enables us to acquire the optimal model for the relationship between DBS parameters and brain response, while simultaneously reducing the number of experiments required. **Methods:** We utilized a computational model of Parkinson's disease to generate synthetic data. By sweeping through various parameters such as subthalamic nucleus DBS amplitude, frequency, and pulse width, we estimated the power of globus pallidus internus (GPi) beta (13-30 Hz) for each DBS parameter. This process yielded a total of 200 distinct samples. Subsequently, we randomly allocated 80% of this data for pool training, reserving the remaining 20% as unseen test data. For establishing the link between DBS parameters and GPi beta power, we employed linear regression and non-linear regression models. To train these models, we initially used three training samples, employing both the RS and AL approaches. Following an iterative process, we added one training sample at a time to both models based on the AL and RS approaches until we had a total of 20 training samples. At each iteration, we assessed the performance of both models on the unseen test data and calculated the root mean squared error (RMSE). We repeated this entire process 1000 times. **Results:** The mean RMSE for the AL and RS approaches was 0.043 and 0.039, respectively. The results demonstrated that the AL-based model exhibited superior performance compared to the RS-based model, as evidenced by significantly fewer errors on the unseen dataset. This difference was statistically significant, as determined by a two-sample t-test ($p = 2.33e-07$, $N = 1000$). **Conclusion:** Our study confirms the superiority of our AL approach over the existing method in establishing a personalized connection between DBS parameters and brain response, all while reducing the duration of the experimental procedure.

Keywords: Deep brain stimulation, active learning, optimal sampling, machine learning

---

Abstract ID: 1125

Study Transcriptional Regulation of Toxoplasma gondii During Bradyzoite Development using iPSC Cells as Models

Forouzandeh Farsaei[1], Thomas Keller[1,] Li Min Ting[1], Jiajia Yang[2], Maria Burgos Angulo[2], Thomas V McDonald[2,3], Kami Kim[1]

[1] Department of Internal Medicine, Division of Infectious Disease and International Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, USA;
[2]Department of Molecular Pharmacology and Physiology, Morsani College of Medicine, University of South Florida, Tampa, FL, USA;
[3]Heart Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA.

Abstract
Toxoplasma gondii, an obligate intracellular parasite, is the causative agent of toxoplasmosis. It is the most common eukaryotic pathogen, and it affects approximately one-third of the global population which in some geographical areas reaches more than 80% of the population. T. gondii's complicated life cycle includes two infectious phases: the replicative (tachyzoite) and latent tissue cyst (bradyzoite) that cause primary and chronic infection, respectively. Primary infection is usually followed by the establishment of latent infection through the formation of intracellular tissue cysts, which tend to form in the brain, eyes, heart, and kidney. This transformation allows the parasite to persist within the host, where immunosuppression can lead to the reactivation of chronic infection. Reactivation of tissue cysts in neuron and muscle cells causes tissue destruction, leading to devastating diseases such as encephalitis, the main cause of death after reactivation in patients with HIV or other immunocompromise. Currently, there is no available drug to treat or prevent the development of these tissue cysts. Therefore, understanding the mechanism and regulatory pathways that lead to the formation and reactivation of cysts is essential to develop effective treatments. A major obstacle in improving our understanding of T. gondii infection is the lack of relevant human models to study cyst formation. To overcome this obstacle, our lab used human induced pluripotent stem cell-derived (iPSC) cardiomyocytes (iCM), cardiac fibroblasts (iCF), and neurons (iNE) as biologically relevant models to study T. gondii infection progression and interaction with the human hosts in vitro. iPSC cells derived from a healthy human donor were differentiated into iCM, iCF, and iNE. They were then infected with T. gondii type I/III EGS reporter strain that expresses specific tachyzoite and bradyzoite markers and FACS analysis and microscopy were performed. We show in vitro human iPSC models can support T. gondii growth and development. RNA sequencing results demonstrated significant differences in the transcriptional profile of T. gondii grown in different iPSC cells. Bradyzoites form spontaneously in iCM and iN and bradyzoite-specific markers were significantly expressed in iCM and iNE following infection. These findings indicate that our in vitro generated human iPSCs provide a novel model to study T. gondii gene expression patterns in different host cell types, and enhance our comprehension of stage differentiation in T. gondii.

Keywords: Toxoplasma gondii; bradyzoite; iPSC; chronic infection; cyst; latent infection

---

Abstract ID: 1137
Drug-Drug Interaction Prediction in Diabetes Mellitus

Rashini Maduka[1], Rupika Wijesinghe[1*], Ruvan Weerasinghe[1]

[1] School of Computing, University of Colombo, Sri Lanka
[*]Corresponding author

Abstract

Drug-drug interactions (DDIs) can happen when two or more drugs are taken together. Today DDIs have become a serious health issue due to adverse drug effects. In vivo and in vitro methods for identifying DDIs are time consuming and costly. Therefore, in silico-based approaches are preferred in DDI identification. Most of machine learning models for DDI prediction are used chemical and biological drug properties as features. However, some drug features are not available and costly to extract. Therefore, it is better to make automatic feature engineering. Furthermore, people who have diabetes already suffer from other diseases and take more than one medicine together. Then adverse drug effects may happen to diabetic patients and make unpleasant reactions in the body. In this study we present a model with a graph convolutional auto encoder and a graph decoder using a dataset from DrugBank version 5.1.3. Main objective of the model is to identify unknown interactions between antidiabetic drugs and the drugs taken by diabetic patients for other diseases. We considered automatic feature engineering and used Known DDIs only as the input for the model. Our model has achieved 0.86 in AUC and 0.86 in AP. Using in-silico based methods, drug-drug interactions can be predicted with less time and cost. It will help to reduce adverse drug effects.

Index Terms: Drug-drug interaction prediction, Graph Convolution Networks, Graph embedding

---

Abstract ID: 1148
Machine Learning Analysis for Studying Aging-Associated Hearing Loss

Safa Shubbar[1], Qiang Guan[1], John W. Hawks[2], Jianxin Bao[3]

[1]Department of Computer Science, Kent State University, Kent, OH, USA;

[2] Kent State University, Kent, OH, USA;

[3] Northeast Ohio Medical University (NEOMED), Ravenna, Ohio, USA

Abstract
Presbycusis, or age-related hearing loss (ARHL), is a condition marked by a gradual decline in auditory sensitivity, involving the loss of sensory cells and central processing functions associated with aging. The key features of ARHL include difficulty in hearing high-pitched sounds, reduced ability to understand speech in noisy or echoey environments, trouble with detecting quick changes in speech, and impaired ability to locate the source of sounds. According to the World Health Organization (WHO) approximately 30% of people over 60 years of age have hearing loss. By 2050, an estimated 2.45 billion people worldwide will experience hearing loss, which is a 56.1% increase from 2019. We did data-driven cluster analysis (k-means clustering) in total dataset (subjects with hearing loss audiograms (n = 733 and 15 features)), dataset 1 (hearing loss audiograms) and dataset 2 (Audiometric shape parameters) to verify that the cluster structure described for each dataset was reproducible

Keywords: Hearing Loss, Aging, Age related, Unsupervised Machine Learning.

Abstract ID: 1170
Quantifying the Growth of Glioblastoma Tumors Using Multimodal MRI Brain Images

Anisha Das, Shengxian Ding, Rongjie Liu and Chao Huang

Department of Statistics, Florida State University, Tallahassee, Florida, USA

Abstract
Predicting the eventual volume of malignant or benign cells that might proliferate from a given tumor, can help in its early detection and subsequently the desired medical procedures can be applied to stop the proliferation of such cells thus preventing their migration to other organs. In this work, a new formulation of the detection problem has been done using Bayesian technique for finding the eventual volume of such cells expected to proliferate from a Glioblastoma (GBM) tumor. The location of the tumor has been determined using parallel image segmentation algorithm. Once the location is determined, we find out how many cells can proliferate from this tumor until its survival time. For this, we start off by finding the likelihood of the eventual number of tumor cells that can take birth in a given time frame. We choose a certain prior subject to logic and our data set structure; and finally obtain our posterior distribution. Using the posterior mean, we obtain our desired eventual volume of tumor cells that need to be predicted. We also determine the corresponding probability that no tumor cell goes undetected when we find the ultimate eventual volume. The model so developed gives excellent results on our dataset. It is expected that the model will work on any dataset where the changes are not measured with regards time. We extend the model and run a Bayesian regression to incorporate other radiomic features of the tumor and discover that their inclusion enhances the chances of no tumor cells remaining undetected. We have mainly focused on detection of malignant cells, but the same model can be used for detecting both malignant and benign cells.

Keywords: Glioblastoma (GBM), malignant cells, proliferation, Bayesian technique, posterior mean

---

Abstract ID: 1177
Enhancing DNA Sequence Matching and Ranking through Deep Learning-Based Alignment-Free Model

Sumarga K. Sah Tyagi[1], Minh Pham[1], Yicheng Tu[1]

[1]Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA.

Abstract: Refer to Flash Talk Session

Abstract ID: 1200
Uncovering ancestry-specific differences in genetically driven transcriptomic dysregulation in schizophrenia and bipolar disorder

Deepika Mathur[1-5], Sanan Venkatesh[1-5], Tim Bigdeli[8-11], Karen Therrien [1-6], Prashant NM, Gabriel Hoffman[1-5], Jaroslav Bendl[1-5], Donghoon Lee[1-5], Biao Zeng[1-5], Aram Hong[1-5], Clara Casey[1-5], Marcela Alvia[1-5], Zhiping Shao[1-5], Stathis Argyriou[1-5], Pavan Auluck[7], David Bennett[13], Stefano Marenco[7], Vahram Haroutunian[1-5], Ayman Fanous[12], Philip Harvey[10,11], John Fullard[1-5], Georgios Voloudakis[1-5] *, Panos Roussos[1-5,12] *

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[2]Center for Disease Neurogenomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[3]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[4]Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[5]Mental Illness Research, Education, and Clinical Center, James J. Peters VA Medical Center, Bronx, NY, USA.
[6]Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[7]Human Brain Collection Core, National Institute of Mental Health-Intramural Research Program, Bethesda, MD, USA.
[8]VA New York Harbor Healthcare System, Brooklyn, NY, USA.
[9]Institute for Genomics in Health, SUNY Downstate Health Sciences University, Brooklyn, NY, USA.
[10]Department of Psychiatry and Behavioral Sciences, SUNY Downstate Health Sciences University, Brooklyn, NY, USA.
[11]Department of Epidemiology and Biostatistics, School of Public Health, SUNY Downstate Health Sciences University, Brooklyn, NY, USA.
[12]Department of Psychiatry, University of Arizona College of Medicine-Phoenix, Phoenix, AZ, USA.
[13]Carl T. Hayden Veterans Affairs Medical Center, Phoenix, AZ, USA.
[10]Bruce W. Carter Miami Veterans Affairs Medical Center, Miami, Miami, FL, USA.
[11]University of Miami Miller School of Medicine, Miami, FL, USA.
[12]Center for Dementia Research, Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA.
[13]Rush Alzheimer's Disease Center, Rush University Medical Center
[*]co-senior

Abstract
Genome-wide association analyses (GWASs) identify genetic variants associated with disease; however, translating genetic findings to downstream applications is challenging since most variants reside in non-coding regions. Transcriptome-wide association studies (TWASs) leverage tissue-specific molecular profiling data (for the construction of transcriptomic imputation models) to bridge this gap by converting genetic risk variation to transcript-level dysregulation. However, the current status quo as it pertains to the use of TWAS for gene discovery is limited to the usage of GWAS and transcriptomic imputation models of European (EUR) ancestry. Here, we explore the shared and unique components of genetically driven transcriptomic dysregulation in specific brain cell subtypes by leveraging single-nucleus RNA sequencing (snRNAseq) between EUR and African (AFR) ancestry for schizophrenia (SCZ) and bipolar disorder (BD). We leverage molecular profiling data (genotypes and snRNAseq expression) from Neuropsychiatric Symptoms in Alzheimer's Disease (NPS-AD) project's dorsolateral prefrontal cortex brain samples comprising of 880 EUR and 253 AFR individuals to train ancestry-specific transcriptomic imputation models. SnRNAseq reveals the genes with altered expression in specific cell subtypes in SCZ and BD. An additional EUR imputation model is generated using 253 individuals matched in terms of sample size, age, sex, and disease status to the individuals making up the AFR model (to compare power). We assess the trans-ancestry predictive performance of the EUR transcriptomic imputation model to predict observed AFR expression and vice versa. We then perform cis-ancestry summary-level single nucleus transcriptome-wide association studies (snTWAS) for EUR using the latest publicly available SCZ (65,205 cases/87,919

controls) and BD (41,917 cases/371,549 controls) GWASs from the Psychiatric Genomics Consortium 3 (PGC3). For the cis-ancestry AFR snTWAS, we leverage GWAS summary statistics from the Genomic Psychiatry Cohort (GPC), Million Veteran Program (MVP) and CSP #572 for SCZ (7,697 cases/ 8,498 controls) and BD (3,027 cases/ 7,988 controls). Finally, using the cis-ancestry snTWASs as ground truth, we (1) explore shared and unique gene and pathway dysregulation across cell subtypes in these disorders and (2) perform trans-ancestry snTWASs to evaluate the loss of power when using models of different ancestry. We find greater convergence of gene expression dysregulation at the gene and pathway level in specific cell subtypes in SCZ than BD between ancestries. Trans-ancestry snTWASs exhibit significant loss of power but overall maintain moderate correlation in top genes.

Keywords: snTWAS, snRNAseq, GWAS, ancestry-specific, schizophrenia, bipolar disorder

---

Abstract ID: 1206

Spatial single-cell transcriptome analysis reveals cellular and molecular heterogeneity, signatures, and pathological trajectories of fatal SARS-CoV-2 lung infection.

Arun Das[1,2], Wen Meng[1,3], Zhentao Liu[1,4], Md Musaddaqul Hasib[1,2], Hugh Galloway[1,4], Suzane Ramos da Silva[1,3], Luping Chen[1,3], Gabriel L Sica[5], Alberto Paniz-Mondolfi[6], Clare Bryce[6], Zachary Grimes[6], Emilia Mia Sordillo[6], Carlos Cordon-Cardo[6], Karla Paniagua Rivera[7], Mario Flores[7], Yu-Chiao Chiu[2,8], Yufei Huang[1,2,4*], and Shou-Jiang Gao[1,3*]

[1]Cancer Virology Program, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[2]Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[3]Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[4]Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, USA
[5]Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[6]Department of Pathology, Molecular and Cell-Based Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA
[7]Department of Electrical and Computer Engineering, KLESSE School of Engineering and Integrated Design, University of Texas at San Antonio, San Antonio, TX, USA
[8]Cancer Therapeutics Program, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
*Corresponding authors. Emails: yuh119@pitt.edu; gaos8@upmc.edu.

Abstract

Despite numerous research incorporating single-cell RNA sequencing (scRNA-seq), spatial transcriptomics, and morphology imaging over the last three years, the pathology, and the underlying molecular mechanism of coronavirus disease 2019 (COVID-19) remains poorly defined. Taking advantage of single-cell spatially resolved transcriptomics technologies, in this study, we developed a spatial single-cell transcriptome analysis (SSCTA) workflow to examine postmortem COVID-19 lung tissues. We identified 11 parenchymal and 7 immune cell types (18 major cell types), all of which were infected by the SARS-CoV-

2 virus. Compared to non-COVID control lung samples, the COVID-19 samples saw major reductions in alveolar cells (ACs) and an increase in innate and adaptive immune cells. Differential expression analysis of the infected samples versus non-COVID control revealed 11 upregulated and 10 downregulated genes suggesting traits of pulmonary fibrosis, disruption of vascular barriers, complement activation and inflammation, and dysregulation associated with pulmonary fibrosis. Spatial analysis using bi-variate Moran's I revealed locally distinct high-infection regions correlated with high-density (HIHD) that expressed high levels of SARS-CoV-2 entry-related factors, including ACE2, FURIN, TMPRSS2, and NRP1, which are colocalized with Organizing Pneumonia (OP), lymphocytic and immune infiltration. We found that HIHD regions had a significantly higher number of ACs and fibroblasts by decreased vascular endothelial cells and epithelial cells which reflects the tissue damage and wound healing process. The neighborhood cell type composition (NCTC) was generated by counting unique cell types in the nearest neighbors of each cell. Sparse non-negative matrix factorization (SNMF) analysis of NCTC features identified seven spatial pathology signatures that captured the structure and immune niches in COVID-19 tissues. Furthermore, we found two distinct pathological trajectories based on the immune niche signatures where Trajectory A likely reflected the complication of microbial infections marked by an increase of NK cells and granulocytes and Trajectory B accounted for the increased immune infiltration marked by a majority of cells from the HIHD and OP regions. The OP regions were marked by an increased composition of fibroblasts with high expressions of COL1A1 and COL1A2. We identified similar cell populations of myofibroblasts in scRNA-seq examinations of COVID-19 and idiopathic pulmonary fibrosis lung tissues. The activation of IL6-STAT3 and TGF-β-SMAD2/3 pathways in these cells was revealed using Immunofluorescence staining, which likely mediated the upregulation of COL1A1 and COL1A2, and excessive fibrosis in the lung tissues. Together, this study reveals the complex cellular and molecular signatures of COVID-19 lungs, revealing their complex spatial cellular heterogeneity, organization, and interactions.

Keywords: spatial single-cell transcriptome analysis; COVID-19; neighborhood cell type composition analysis; trajectory inference; organizing pneumonia; fibrosis.

Abstract ID: 1231

Disentangling accelerated cognitive decline in Alzheimer's disease from the normal aging process and identifying its genetic underpinnings: A deep learning approach using neuroimaging

Yulin Dai[1], Yu-Chun Hsu[2], Astrid M Manuel[1], Andi Liu[1,3], Brisa S Fernandes[1], Jingchun Chen[4], Xiaoyang Li[1,3], Nitesh Enduru[1], Kai Zhang[2], Xiaoqian Jiang[2], Zhongming Zhao[1,5,6]
Detailed Affiliations

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2]Center for Secure Artificial Intelligence for Healthcare, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA
[4]Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, Las Vegas, NV 89154, USA
[5]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[6]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA

Abstract
Alzheimer's disease (AD) has surfaced as one of the leading health problems of the 21st century, with enormous societal and economic costs; currently, AD affects over 6 million people in the US alone. AD is a disease characterized by an accelerated cognitive decline compared to what would be expected by the normal aging process. AD has no cure or prevention, partly due to the absence of a coherent and actionable model to differentiate its progression in the earlier stages from normal aging, leading to a lack of effective early intervention and prevention treatments. To assess how a particular person's cognitive decline departs from what would be expected by normal aging, we need to measure differences between chronological age and cognitive decline at different time points. To achieve this aim, we developed a deep-learning framework to extract fine-grained information from the longitudinal structural T1-weighted magnetic resonance imaging (T1w-MRI) data in Alzheimer's Disease Neuroimaging Initiative (ADNI). Specifically, we used 3D Residual Network (ResNet) and Siamese network to extract and learn the sequentially paired neuroimaging and cognitive score data across 9,680 data points from 1,313 individuals. We trained our model on 414 Cognitive normal individuals and predicted their cognitive assessment on these 1,313 individuals. Conditional to the normal aging predicted neuroimaging features, we conducted a genome-wide association study (GWAS) of the cognitive decline slope with paired genotyping data to further reveal the genetic basis of AD disease progression and AD-related accelerated cognitive decline. We successfully identified two genome-wide significant loci, the well-known APOE and one locus on Chr11, a newly identified locus never reported in previous AD GWAS. The single-cell colocalization and fine mapping analysis linked this new locus to the neural epidermal growth factor (EGF)-like repeats (NELL1) gene, which expressed protein kinase C-binding protein NELL1, plays a role in control of cell growth and differentiation. The cell-type-specific enrichment analysis and functional enrichment of GWAS signals highlighted the synapse, microglia, and immune-response pathways. Lastly, we observed a positive correlation between the cognitive decline slope and previous AD GWAS, while a negative correlation was found with intelligence and educational attainment GWAS. In conclusion, our deep learning model corrected the noise in cognitive assessments and identified a novel implicated gene. Our approach has the potential to disentangle accelerated cognitive decline from the normal aging process and to determine its related genetic factors, propitiating opportunities for early intervention.

Keywords: Alzheimer's disease, longitudinal T1-weighted magnetic resonance imaging, Deep Learning, Cognitive Decline, Genome-wide association study, single-cell

---

Abstract ID: 1234
Bioinformatics and machine learning based identification of potential oxidative stress and glucose metabolism diagnostic Biomarkers in Alzheimer disease.

Sidra Aslam[1], Fatima Noor[2], Thomas G. Beach[1], Geidy E. Serrano[1]

[1]Banner Sun Health Research Institute, Sun City, AZ, USA
[2] Government College University Faisalabad, Pakistan

Abstract: Refer to Flash Talk Session

Abstract ID: 1235

Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning

Yuzhou Chang[1,*], Fei He[2,*], Juexin Wang[3,*], Shuo Chen[4], Jingyi Li[2], Jixin Liu[5], Yang Yu[2], Li Su[3], Anjun Ma[1], Carter Allen[1], Yu Lin[6], Shaoli Sun[7], Bingqiang Liu[5], José Javier Otero[8], Dongjun Chung[1,9], Hongjun Fu[4], Zihai Li[9,$], Dong Xu[3,$], Qin Ma[1,9,$]

[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA
[2] School of Information Science and Technology, Northeast Normal University, Changchun, Jilin 130117, China
[3] Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[4] Department of Neuroscience, The Ohio State University, Columbus, OH 43210, USA
[5] School of Mathematics, Shandong University, Jinan 250100, China
[6] School of Artificial Intelligence, Jilin University, Changchun 130012, China
[7] Department of Pathology, The Ohio State University, Columbus, OH 43210, USA
[8] Departments of Neuroscience, Pathology, Neuropathology, The Ohio State University, Columbus, OH 43210, USA
[9] The Pelotonia Institute for Immuno-oncology, The Ohio State University Comprehensive Cancer Center, Columbus, OH 43210, USA

Abstract

Spatially resolved transcriptomics provides a new way to define spatial contexts and understand pathogenesis of complex human diseases. Although some computational frameworks can characterize spatial context via various clustering methods, the detailed spatial architectures and functional zonation often cannot be revealed and localized due to the limited capacities of associating spatial information. We present RESEPT, a deep-learning framework for characterizing and visualizing tissue architecture from spatially resolved transcriptomics. Given inputs as gene expression or RNA velocity, RESEPT learns a three-dimensional embedding with a spatial retained graph neural network from the spatial transcriptomics. The embedding is then visualized by mapping into color channels in an RGB image and segmented with a supervised convolutional neural network model. Based on a benchmark of 10x Genomics Visium spatial transcriptomics datasets on the human and mouse cortex, RESEPT infers and visualizes the tissue architecture accurately. It is noteworthy that, for the in-house AD samples, RESEPT can localize cortex layers and cell types based on pre-defined region- or cell-type-specific genes and furthermore provide critical insights into the identification of amyloid-beta plaques in Alzheimer's disease. Interestingly, in a glioblastoma sample analysis, RESEPT distinguishes tumor-enriched, non-tumor, and regions of neuropil with infiltrating tumor cells in support of clinical and prognostic cancer applications.

Keywords: Deep learning, dimension reduction, spatial transcriptomics, tissue heterogeneity, tissue architecture, visualization.

Abstract ID: 1249
TSSr: an R package for comprehensive analyses of TSS sequencing data

Zhaolian Lu[1], Keenan Berry[2], Zhenbin Hu[1], Yu Zhan[1], Tae-Hyuk Ahn[2,3], Zhenguo Lin[1,2]

[1]Department of Biology, Saint Louis University, St. Louis, MO 63103, USA;
[2]Program of Bioinformatics and Computational Biology, Saint Louis University, St. Louis, MO 63103, USA; [3]Department of Computer Sciences, Saint Louis University, St. Louis, MO 63103, USA

Abstract: Refer to Additional Flash Talks

Abstract ID: 1260
Feasibility of a 3D Convolutional Neural Network for the Diagnosis of Alzheimer's Disease using Brain PET Scans

Troy Zhang[1], Yan Guo[2], Yang Mi[2]

[1] Interlake High School, Bellevue WA 98008, USA
[2] University of Miami, Miami FL 33136, USA

Abstract: Refer to Session "Artificial Intelligence on Big Data: Promise for Early-stage Trainees".

Abstract ID: 1278
An integrative study to identify the link between dysregulated intercellular signaling and genetic variants in Alzheimer's disease

Andi Liu[1,2], Xiaoyang Li[2,3], Brisa S Fernandes[2], Yulin Dai[2], Zhongming Zhao[1,2,4,*]

[1]Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA;
[2]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA;
[3]Biostatistics & Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA;
[4]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract: Refer to Additional Flash Tallks

Abstract ID: 1313

Atypical joint density distribution of resting state dynamic spatial brain network pairs in Schizophrenia

Krishna Pusuluri[1], Armin Iraji[1], Vince D Calhoun[1]
[1]Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, and Emory University, Atlanta, GA.

Abstract:
Most resting state functional magnetic resonance imaging (rsfMRI) studies in schizophrenia (SZ) focus on spatially static networks over the course of a scan. Recent studies addressed the spatial expansion or shrinkage (Iraji et al. 2020) and variations in coupling between spatially dynamic networks via a model of the spatial chronnectome (Iraji et al. 2019). In this work, we investigated spatially dynamic brain networks, their voxel-wise changes over time, and the joint density distributions of pairs of networks using 2D histograms, clustered across time windows.
2D histograms allow for the comparison of two networks, counting the occurrence of various combinations of voxel-level intensities/activities. rsfMRI data for 508 subjects with 315 controls (CN) and 193 SZ patients were obtained from three datasets – FBIRN (Damaraju et al. 2014), COBRE (Aine et al. 2017), and MPRC (Adhikari et al. 2019) – and preprocessed using SPM12 toolbox as described in (Iraji et al. 2022). We performed spatial independent component analysis (sICA) at the group level using the GIFT software package (Iraji et al. 2021) and identified 14 large-scale brain networks. The prior networks from group-level analysis were used as a reference for a spatially constrained ICA (integrated in GIFT as Multivariate Objective
Optimization ICA with Reference) (Du et al. 2013) for each subject across sliding windows to ensure the correspondence of brain networks across subjects and time windows (window size is 30 times TR, the repetition time). For each brain network pair (see Fig.1), 2D histograms were computed for z-scored voxel-level activity at each window for each subject and clustered across all the subjects and windows using k-means algorithm (after subtracting the overall mean 2D histogram from the window-level data). Clustering results reveal changes in the joint density heatmaps for the network pair across subjects and time windows. We identified several clusters of 2D histograms that show significant group differences (with 2-sample t-tests) in subject-wise dwelling time (defined as the number of windows per subject that fall within the cluster, divided by the total number of windows), revealing atypical joint density distributions of dynamic spatial brain network pairs in Schizophrenia. This work underscores the importance of studying spatiotemporally dynamic behaviors within/across brain networks and could lead to the development of novel biomarkers for brain function and dysfunction.

Keywords: Functional MRI, Resting state, Psychiatric Disorders, Schizophrenia, Dynamic spatial brain networks

---

Abstract ID: 1322
Discovery of Small Molecules that Induce Yamanaka Factors in Neuronal Cells for the Treatment of Alzheimer's Disease: A Chemogenomic Approach

Kun-Hyung Roh1, Charles V. Mobbs2
1Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, NY, USA; 2Department of Neuroscience, Icahn School of Medicine at Mount Sinai, USA.

Abstract
While partial cellular reprogramming has previously shown to reverse aging phenotypes and regenerate healthy cells in diseased conditions, transgenic approaches confer serious safety concerns for future applications in regenerative medicine. Accordingly, small molecule-induced cellular reprogramming may be an advantageous strategy to reverse aging and age-related diseases. To develop novel drugs and explore significant targets to treat aging and age-related diseases, we used a computational approach to find drugs that transcriptomically induce three of the Yamanaka Factors. We virtually screened the Connectivity Map database to discover drugs that induce OCT4, SOX2, and KLF4, and analyzed them in order of integrated scores indicating transcriptional induction. Among the most promising drugs were histone methyltransferase inhibitors, HDAC inhibitors, and PKC inhibitors, along with other kinase inhibitors. Interestingly, some of the drugs were already previously demonstrated for its neuroprotective properties against Alzheimer's Disease pathology, although their regenerative potential was not defined in the studies. Subsequent chemogenomic analysis revealed that the common targets of the effective drugs are DPYSL2, IGFBP2, ELOVL6, CSRP1, and CSRP2, which suggests that inhibition of such genes in neural cells can potentially induce reprogramming-like signature. Some drugs identified were known to cross the blood-brain barrier, suggesting that repurposing of such drugs would be ideal for the treatment of Alzheimer's Disease. Such computational screening methods have identified potential drugs and genes that may reprogram specific cells, which may be further studied for developing therapeutic methods that protect against age-related diseases in general. Future studies will focus on experimental validation of the geroprotective compounds and develop a machine learning model that predicts such properties.

Keywords: Transcriptomics, Drug Discovery, Machine Learning, Alzheimer's Disease, Neurodegeneration, Aging

---

Abstract ID: 1329
Do Single-cell Hi-C Data Follow a Power Law Distribution?

Bin Zhao[1,2], Patrick Shen[3], Lu Liu[2]

[1]Department of Statistics, North Dakota State University, Fargo, ND, USA;
[2]Department of Computer Sciences, North Dakota State University, Fargo, ND, USA; [3]Davies High School, Fargo, ND, USA.

Abstract
Power law distributions are prevalent in many natural and social systems and can reveal underlying mechanisms and structures in complex systems. Single-cell Hi-C datasets have become increasingly popular in biology due to their ability to capture three-dimensional chromatin organization at the single-cell level. While the presence of power law distributions has been established in bulk Hi-C data, it remains unexplored whether single-cell Hi-C data exhibit the same characteristic pattern. In this study, we analyzed power law distributions in single-cell Hi-C datasets ranging from base to 1Mb resolution, using two testing methods: hypothesis testing and likelihood ratio testing. Our results demonstrate that the majority of single-cell Hi-C data follow a power law distribution, indicating that power law behavior is a fundamental property of

biological systems. Our finding can be applied to developing new computational methods for single-cell Hi-C data.

Keywords: Power law distribution; single-cell; Hi-C, Hypothesis test; Likelihood ratio test

---

Abstract ID: 1343
The Nonlinear Brain: Estimating Networks from Explicitly Nonlinear Functional Connectivity

Spencer Kinsey[1], Armin Iraji[1], Katarzyna Kazimierczak[2], Jiayu Chen[1], Sara Motlaghian[1], Karsten Specht[2], Tulay Adali[3], and Vince D. Calhoun[1]

[1]Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Atlanta, GA, USA
[2]Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway
[3]Department of CSEE, University of Maryland, Baltimore, MD, USA

Abstract

Introduction: Functional magnetic resonance imaging (fMRI) studies often estimate brain networks from temporal relationships between BOLD signals, which is commonly called functional connectivity (FC). Independent component analysis (ICA) is one method that can be applied to fMRI time series to obtain maps of brain regions (intrinsic connectivity networks (ICNs)) that exhibit similar temporal fluctuations [1]. ICA may also be applied to FC matrices constructed from second-order statistics such as Pearson correlation [2]. However, both approaches are typically designed to identify ICNs whose elements covary similarly, reflecting linear FC. Although insights from linear FC have contributed extensively to our understanding of brain function, many brain processes have nonlinear aspects [3]. Nonlinear variants of ICA exist [4], but ICA is not commonly used to estimate networks from explicitly nonlinear patterns. To extract the potentially valuable information contained in these patterns, we propose a novel approach to estimate brain networks, for the first time, from explicitly nonlinear whole-brain FC information. **Materials and Methods:** We analyzed a multi-site resting-state fMRI data set comprised of 508 subjects (315 controls and 193 individuals diagnosed with schizophrenia) [5]. Each subject's data were preprocessed to correct for motion-related artifacts and scanner distortion, and variance normalized. After calculating both whole-brain covariance (LIN-wFC) and distance correlation (NL-wFC) [6], we removed LIN-wFC information from NL-wFC with a regression-based method to obtain the explicitly nonlinear whole-brain FC (ENL-wFC) for every subject. Applying twenty model order ICA to LIN-wFC and ENL-wFC allowed us to obtain large-scale ICNs and compare them. **Results**: Group ICA of fMRI toolbox (GIFT) ICASSO [7] stability suggests that ENL-ICN estimation is more stable than LIN-ICN estimation. Results show that most networks calculated from LIN-wFC are also represented in ENL-wFC, with ten ICNs exhibiting a spatial correlation value exceeding 0.80. Additionally, we identified a unique LIN-ICN as well as a unique ENL-ICN, indicating the potential of our method to reveal brain networks that would otherwise be hidden. Voxel-wise paired samples t-testing between spatially similar ENL and LIN-ICNs revealed that ENL-ICN weight was greater within core regions. Moreover, a two-sample t-testing approach between ICNs derived from controls and those derived from schizophrenia patients revealed that some ENL-ICN comparisons were more sensitive to group differences, demonstrating the translational value of our approach. In summary, our

findings indicate that incorporating nonlinear patterns within the scope of fMRI FC analysis may shed light on brain function and differentiate clinical cohorts.

Keywords: fMRI ICA nonlinear connectivity network schizophrenia

---

Abstract ID: 1350
Identifying relationships between cellular topology and gene expression in spatial transcriptomics of breast cancer tissues

Isabella Wu[1], Chen Li[2], Wentao Huang[2], Debolina Chatterjee[3], Jie Zhang[3], Chao Chen[2*], Travis S. Johnson[3*]

[1]Choate Rosemary Hall High School, Wallingford, Connecticut; [2]Stony Brook University, Stony Brook, New York; [3]Indiana University School of Medicine, Indianapolis, Indiana
[*]Corresponding authors: chao.chen.1@stonybrook.edu and johnstrs@iu.edu

Abstract: Refer to Session "Artificial Intelligence on Big Data: Promise for Early-stage Trainees".

Abstract ID: 1352
Comprehensive Investigation of Active Learning Strategies for Anti-Cancer Drug Response Prediction

Priyanka Vasanthakumari[1], Yitan Zhu[1], Thomas Brettin[2], Alexander Partin[1], Maulik Shukla[1], and Rick L. Stevens[1,3]

[1] Division of Data Science and Learning, Argonne National Laboratory, Lemont, IL, USA
[2] Computing, Environment and Life Sciences Directorate, Argonne National Laboratory, Lemont IL, USA
[3] Department of Computer Science, The University of Chicago, Chicago, IL, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1353
MalariaSED: a deep learning framework to decipher the regulatory contributions of noncoding variants in malaria parasites

Chengqi Wang[1]*, Yibo Dong[1], Jenna Oberstaller[1], Chang Li[1], Min Zhang[1], Justin Gibbons[1], Camilla Valente Pires[1], Lei Zhu[4], Rays H.Y. Jiang[1], Kami Kim[2], Jun Miao[2], Thomas D. Otto[3], Liwang Cui[2], John H. Adams[1], Xiaoming Liu[1]

[1]Center for Global Health and Infectious Diseases Research and USF Genomics Program, College of Public Health, University of South Florida, Tampa, FL, USA.

[2]Department of Internal Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, USA
[3]School of Infection & Immunity, MVLS, University of Glasgow, Glasgow, UK.
[4]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

Abstract: Refer to Flash Talk Session

Abstract ID: 1372
Machine learning generated insights from Swiss COVID-19 patients lung proteomics data

Paola Martinez Murillo[1], Pierre-Yves Mantel[1,2], Swamy R. Adapa[3] and Rays H.Y. Jiang[3]

[1] Christine Kühne – Center for Allergy Research and Education, Davos, Switzerland
[2] Department of Oncology, Microbiology, and Immunology, University of Fribourg, Fribourg, Switzerland.
[3] USF genomics, Global Health Infectious Disease Research Center, College of Public Health, University of South Florida, Tampa, FL, USA

Abstract
We present an ongoing project employing ensemble Machine Learning (ML) to uncover unique mechanisms of COVID-19 pathogenesis distinct from other lung infections, leveraging high-quality clinical data from Swiss hospitals. By isolating extracellular vesicles (EVs) from the lungs of COVID-19 patients via Broncho alveolar lavage (BAL) and conducting mass spectrometry-based proteomics analysis, we characterize the COVID-19 lung proteome. Our initial study, encompassing 50 COVID-19 and 50 non-COVID control samples, expands to include plasma EVs for comparison with BAL EVs in larger patient cohorts. We develop a machine learning model utilizing proteomic markers to predict COVID-19 infection and prioritize human markers. Our findings encompass the overall characterization of the COVID-19 lung proteome, identification of COVID-19-specific cellular pathways, and the development of a predictive machine learning model. Supported by comprehensive clinical data, our research offers promise in advancing understanding, targeted interventions, and personalized treatment strategies for COVID-19.

Abstract ID: 1394
Cancer Comprehend Annotation – a pipeline for cancer phenotype and clinical extraction

Thanh Duong[1], Phillip Szepietowski[2], Thanh Thieu[1]

[1]Department of Machine Learning, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States
[2]Department of Health Data Services, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States.

Abstract: Refer to Additional Flash Talks

Abstract ID: 1429
Developing an Accurate and Interpretable Risk-Based Model for Lung Cancer Screening

Piyawan Conahan[1], Lary Robinson[2], Haley Tolbert[3], Margaret M Byrne[4], Lee Green[4], Yi Luo[1]

[1]Department of Machine Learning, Moffitt Cancer Center, FL, USA;
[2]Division of Thoracic Oncology (Surgery), Moffitt Cancer Center, FL, USA;
[3]Lung Cancer Screening Program, Moffitt Cancer Center, FL, USA;
[4]Department of Health Outcomes and Behavior, Moffitt Cancer Center, FL, USA.

Abstract: Refer to Flash Talk Session

Abstract ID: 1442
The RNA m^{6}A landscape of mouse oocytes and preimplantation embryos]

Yunhao Wang[1,2,11], Yanjiao Li[3,4,11], Trine Skuland[5,6], Chengjie Zhou[7#], Aifu Li[1,2], Adnan Hashim[3], Ingunn Jermstad[8], Shaista Khan[8], Knut Tomas Dalen[8], Gareth D. Greggains[5], Arne Klungland[3,9]*, John Arne Dahl[3]*, Kin Fai Au[1,2,10]*

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA;
[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA;
[3]Department of Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway;
[4]Department of Molecular Medicine, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway;
[5]Department of Reproductive Medicine, Oslo University Hospital, Oslo, Norway;
[6]Division of Gynaecology and Obstetrics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway;
[7]State Key Laboratory of Reproductive Regulation & Breeding of Grassland Livestock, School of Life Sciences, Inner Mongolia University, Hohhot, Inner Mongolia, China;
[8]Norwegian Transgenic Centre, Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway;
[9]Department of Biosciences, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway;
[10]Biomedical Informatics Shared Resources, The Ohio State University, Columbus, OH, USA;
[11]These authors contributed equally: Yunhao Wang, Yanjiao Li;
[#]Present address: Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA;
[*]Corresponding authors.

Abstract
N^6-methyladenosine (m^{6}A), the most prevalent internal modification in eukaryotic messenger RNA, plays key regulatory roles in many biological processes (e.g., RNA stability, splicing, transport, and translation) and is involved in a variety of physiological processes (e.g., cell differentiation and reprogramming, embryonic development, and stress responses). Despite the significance of m^{6}A, the requirement for large amounts of RNA has hindered m^{6}A profiling in mammalian early embryos. Here, we apply low-input methyl RNA immunoprecipitation and sequencing to map m^{6}A in mouse oocytes and preimplantation

embryos. We define the landscape of m^6A during the maternal-to-zygotic transition, including stage-specifically expressed transcription factors essential for cell fate determination. Both the maternally inherited transcripts to be degraded post fertilization and the zygotically activated genes during zygotic genome activation are widely marked by m^6A. In contrast to m^6A-marked zygotically activated genes, m^6A-marked maternally inherited transcripts have a higher tendency to be targeted by microRNAs. Moreover, RNAs derived from retrotransposons, such as MTA that is maternally expressed and MERVL that is transcriptionally activated at the two-cell stage, are largely marked by m^6A. Our results provide a foundation for future studies exploring the regulatory roles of m^6A in mammalian early embryonic development.

Keywords: N^6-methyladenosine, mouse oocyte and early embryonic development, transcription factor, maternal-to-zygotic transition, microRNA, retrotransposon

---

Abstract ID: 1450

Unveiling Gene Interactions in Alzheimer's Disease by Integrating Genetic and Epigenetic Data with a Network-Based Approach

Keith Sanders[1], Astrid M Manuel[1], Andi Liu[1, 2], Boyan Leng[3], Xiangning Chen[1], Zhongming Zhao[1,2,4]

[1] Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center, Houston, TX 77030, USA; [2] Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center, Houston, TX 77030, USA; [3] Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center, Houston, TX 77030, USA; [4] Human Genetics Center, School of Public Health, The University of Texas Health Science Center, Houston, TX 77030, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1456

Single-cell biological network inference using a heterogeneous graph transformer

Anjun Ma[1,2,*], Xiaoying Wang[3,*], Jingxian Li[3], Cankun Wang[1], Tong Xiao[2], Yuntao Liu[3], Hao Cheng[1], Juexin Wang[4,5], Yang Li[1], Yuzhou Chang[1,2], Jinpu Li[5,6], Duolin Wang[4,5], Yuexu Jiang[4,5], Li Su[5,6], Gang Xin[2], Shaopeng Gu[1], Zihai Li[2], Bingqiang Liu[3,$], Dong Xu[4,5,6, $], Qin Ma[1,2,$]

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA
[2] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
[3] School of Mathematics, Shandong University, Jinan, Shandong, 250100, China
[4] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
[5] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA.
[6] Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

[*] These authors contributed equally
[$] To whom correspondence should be addressed

Abstract
Single-cell multi-omics (scMulti-omics) allows the quantification of multiple modalities simultaneously to capture the intricacy of complex molecular mechanisms and cellular heterogeneity. Existing tools cannot effectively infer the active biological networks in diverse cell types and the response of these networks to external stimuli. Here we present DeepMAPS for biological network inference from scMulti-omics. It models scMulti-omics in a heterogeneous graph and learns relations among cells and genes within both local and global contexts in a robust manner using a multi-head graph transformer. Benchmarking results indicate DeepMAPS performs better than existing tools in cell clustering and biological network construction. It also showcases competitive capability in deriving cell-type-specific biological networks in lung tumor leukocyte CITE-seq data and matched diffuse small lymphocytic lymphoma scRNA-seq and scATAC-seq data. In addition, we deploy a DeepMAPS webserver equipped with multiple functionalities and visualizations to improve the usability and reproducibility of scMulti-omics data analysis.

Keywords: Single-cell multi-omics, heterogeneous graph transformer, multi-head attention mechanism, cell-type-specific biological networks, webserver

---

Abstract ID: 1458
Cross-analysis between P. falciparum Var expression with host immunothrombosis markers to better define pediatric cerebral malaria phenotypes.

Iset Vera[1], Thomas Keller[1], Anne Kessler[2], Visopo Harawa[3,4,7], Wilson L. Mandala[3,4,5], Stephen J. Rogerson[6], Terrie E. Taylor[7,8], Karl B. Seydel[7,8], and Kami Kim[1]

[1] University of South Florida, Tampa, FL, USA
[2] Albert Einstein College of Medicine, Bronx, NY, USA
[3] Malawi-Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi
[4] University of Malawi, College of Medicine, Biomedical Department, Blantyre, Malawi
[5] Academy of Medical Sciences, Malawi University of Science and Technology, Thyolo, Malawi
[6] The University of Melbourne, Melbourne, Australia
[7] Blantyre Malaria Project, Blantyre, Malawi
[8] Michigan State University, East Lansing, MI, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1472
PLXNC1: A Novel Potential Immune-Related Target for Stomach Adenocarcinoma

Zhizhan Ni[1], Chenshen Huang[1], Hongmei Zhao[2], Jinzhe Zhou[1], Muren Hu[1], Qing Chen[3], Bujun Ge[1], Qi Huang[1]

[1] Department of General Surgery, Tongji Hospital, School of Medicine, Tongji University, Shanghai, China;
[2] Department of VIP Clinic, East Hospital, School of Medicine, Tongji University, Shanghai, China;
[3] University, Shanghai, China;3 Department of General Surgery, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China

Abstract
Background: Gastric cancer is associated with tumor microenvironment and chronic inflammation, but the underlying tumor-promoting mechanisms still remain unknown. **Methods:** The ATAC-seq was used to identify genes with chromatin accessibilities in promoter regions. The RNA-seq datasets were performed to identify differentially expressed genes (DEGs). Pearson correlation analysis with the mRNA expression of three families of tumor-related inflammation TFs was used to filter downstream DEGs. Cox univariate survival analysis was performed to identify the prognostic value. The ImmPort database and CIBERSORTx algorithm were used to investigate the regulatory relationship between hub DEGs and immune cells. Immunohistochemistry (IHC) and multidimensional database were performed to verification. **Results:** In this case, we require 2454 genes with chromatin accessibility in promoter regions by ATAC-seq. Based on the gene expression profiles (RNA-seq), we identified 365 genes with chromatin accessibility and differential expression. Combined with the Cox univariate survival analysis, we identified 32survival-related DEGs with chromatin accessibility. According to ImmPort database, CXCL3, PLXNC1, and EDN2 were identified as immune- related genes in STAD. By applying the CIBERSORTx algorithm and Pearson correlation, PLXNC1 was the only gene correlated with various immune cells, significantly associated with M2 macrophages. Furthermore, gene set variation analysis (GSVA) suggests the "hallmark_interferon_gamma_response" pathway was most significantly correlated with PLXNC1. Immunohistochemistry results revealed that PLXNC1 protein level was significantly higher in STAD tissues than in normal tissues ($p < 0.001$). **Conclusion:** PLXNC1, regulated by IRF5, is an immune-related gene that was significantly associated with M2 macrophages and poor outcome in stomach adenocarcinoma.

Keywords: gastric cancer, plexin C1, tumor microenvironment, macrophage, stomach adenocarcinoma, chromatin accessibility.

Abstract ID: 1483
Unveiling the hidden genomic code of insulator elements on a genome-wide scale

Dewan Shrestha[1,2], Qian Qi[2], Sheng Zhou[3], Yong Cheng[1,2,4]

[1]Department of Genetics, Genomics, and Informatics, College of Graduate Health Sciences, The University of Tennessee Health Science Center, Memphis, TN;
[2]Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN;
[3]Experimental Cellular Therapeutics Lab, St. Jude Children's Research Hospital, Memphis, TN;
[4]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN.

Abstract
Insulators are DNA elements that can hinder the functioning of enhancers when positioned between enhancers and promoters. CTCF is the primary transcription factor (TF) associated with insulators in

vertebrates. However, our comprehension of insulators remains limited compared to cis-regulatory elements such as enhancers and promoters. In this study, we have developed a novel three-color fluorescence-based system to quantitatively measure insulator activities within the same chromatin environment. Our framework was applied to analyze over 10,000 DNA sequences occupied by CTCF. Through this high-throughput assay, we were able to confirm the functions of known insulators and discover hundreds of novel insulators. By integrating DNA sequences and evolutionary constraints, TF binding profiles, and epigenomic signatures, we identified several important features that determine the activity of insulators. Notably, our data indicated a significant enrichment of novel 10 bp motifs downstream of core CTCF motifs in CTCF occupancy sites exhibiting strong insulator activities. Additionally, compared to CTCF, the occupancy signals of cohesion showed a higher correlation with insulator activities. Furthermore, we observed a depletion of active histone modification marks such as H3K27ac in insulators, while repressive marks such as H3K27me3 were enriched in insulators. Utilizing machine learning, we constructed models incorporating these different features, which accurately predicted insulator activities (R=0.734 Pearson correlation and R=0.735 Spearman correlation). In summary, our study identified new features of insulator sequences and validated their significance in insulator function. These findings contribute to our understanding of how insulators regulate gene expression.

Keywords: CTCF; Insulator; Gene Therapy; Epigenetics

Abstract ID: 1502
Early Stage or Curable Cancer Diagnoses in Minorities: A Journey of Survivors

Lora Asberry[1], Evelina Sterling[1], Naya Phillips[1]

[1] Kennesaw State University, Kennesaw, GA

Abstract:
Significant technological advancements in the medical field have resulted in more frequent screenings, to increase the chances of catching various types of cancer in its early stages. Screening improvements have helped decrease the cancer mortality rate, increase the cancer survival rate, and inform others who may not be knowledgeable about the benefits of getting an early detection screening; however, there are other implications patients are left with after their diagnosis. Patients depend on relationships with medical personnel to maximize duration and quality of life; however, biases, stigma, access to care/technology, and socioeconomic status may influence patient experiences. The objectives of this project are: to determine whether medical disparities vary between minorities and non-minorities who have early-stage or curable cancer, to analyze the effects of the cancer diagnoses in minorities compared to non-minorities, to assess the different perspectives in minority male vs. female participants, and to demonstrate whether there is a communication barrier between cancer patients and their medical professionals, regarding the health and knowledge of their diagnosis. While various modes of cancer treatment should be accessible to individuals from all walks of life, previous studies have demonstrated that minorities, as well as other individuals from underrepresented communities, do not receive the same level of care in comparison to their non-minority counterparts who also receive cancer treatment. Often times, cancer experiences from underrepresented patients are overlooked; however, it is essential that each of their journeys is properly explored. Due to this disconnect and a lack of extensive research, my study aims to identify the experiences of underrepresented

individuals with early-stage curable cancer diagnoses. By race/ethnicity, gender, age, and cancer diagnosis, participants will discuss the varied and unique experiences of being diagnosed with early-stage or curable cancers. We want our sample to specifically focus on underrepresented minorities and men, who are underrepresented in cancer research studies. This study is analyzed using a modified grounded theories approach of looking for common themes and finding the connections of how they fit together. Expected results presume differences in oncology experiences within minority patients due to possible cultural differences, varying educational statuses, access to care, as well as possible disparities and personal preconceived biases; however, while technological advances are used to detect early staged or curable cancers, other unintended consequences, such as: stigma, fear, and irrational decisions, and impulsivity are prevalent.

Keywords: Cancer, Health Disparities, Minorities, Oncology

Abstract ID: 1503
An integrated strain-level analytic pipeline utilizing longitudinal metagenomic data

Boyan Zhou[1], Chan Wang[1], Gregory Putzel[2], Jiyuan Hu[1], Menghan Liu[3], Fen Wu[4], Yu Chen[4], Alejandro Pironti[2], Huilin Li[1#]

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

Abstract
With the development of sequencing technology and analytic tools, studying within-species variations enhances the understanding of microbial biological processes. However, most existing methods for strain-level analysis do not allow for the simultaneous interrogation of strain proportions and genome-wide variants in longitudinal metagenomic samples. In this study, we introduce LongStrain, an integrated pipeline for the analysis of large-scale metagenomic data from individuals with longitudinal or repeated samples. In LongStrain, we first utilize two efficient tools, Kraken2 and Bowtie2, for the taxonomic classification and alignment of sequencing reads respectively. Then, we propose to jointly model strain proportions and shared haplotypes across samples within individuals, which greatly improves the efficiency and accuracy of strain identification. With extensive simulation studies of a microbial community and single species, we show that LongStrain is superior to three popular reference genome-based methods in variant calling and strain-proportion estimation. Furthermore, we illustrate the potential applications of LongStrain in the real data analysis of The Environmental Determinants of Diabetes in the Young study and a gastric intestinal metaplasia microbiome study. In summary, the proposed analytic pipeline demonstrates marked statistical efficiency over same type of methods and has great potential in understanding the genomic variants and dynamic changes at strain level.

Keywords: Microbiome, longitudinal metagenomic data, strain-level analysis, genomic variants, strain dynamics

---

Abstract ID: 1526
3D genome reveals intratumor heterogeneity in Glioblastoma

Qixuan Wang[1], Juan Wang[1], Qiushi Jin[1], Mark W. Youngblood[1], Lena Ann Stasiak[1], Ye Hou[1], Yu Luan[1], Radhika Mathur[2], Joseph F. Costello[2], Feng Yue[1]

[1]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[2]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1554
Applications of prediction models in the emergency department: a bibliometric analysis

Amirmohammad Shahbandegan[1*], Vijay Mago[1], David W. Savage[2]

[1] Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada
[2] NOSM University, Thunder Bay, Ontario, Canada
[*] Corresponding author, ashahban@lakeheadu.ca

Abstract
The recent advancements in artificial intelligence (AI) is playing a significant role in the growth of research in health/medical sciences, including its sub-domains like emergency medicine. This study aims to provide a comprehensive overview of the existing literature on prediction models in emergency departments (ED) through bibliometric analysis. A literature search was conducted using the Google Scholar platform to find articles related to the ED and prediction models published between 2012 and 2022. The dataset collected contained 423 articles with information such as authors, journals, publication year, and number of citations. Performance analysis was conducted to measure the productivity and quality of AI research in the ED. The most influential articles and journals were selected based on the number of citations received, and the most active researchers and organizations were selected based on the number of articles published. The results of the bibliometric analysis show that machine learning has been widely applied in a variety of applications in EDs and has greatly contributed to the advancement of the field. This study provides insights into the current state of research on AI-based prediction models in EDs and identifies potential directions for future research.

Abstract ID: 1559
Common Genetic Variants are Associated with Plasma and Skin Carotenoid Metabolism in Ethinically Diverse US Populations

Yixing Han[1], Savannah Mwesigwa[2], Melissa N. Laska[3], Stephanie B. Jilcott Pitts[4], Nancy E. Moran[5], Neil A. Hanchard[1]

[1]Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD;
[2]Department of Medical Microbiology, College of Health Sciences, Makerere University, Kampala, Uganda;
[3]Division of Epidemiology & Community Health, School of Public Health, University of Minnesota,
[4]Department of Public Health, East Carolina University, Greenville, NC;
[5]USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX.

Abstract: Refer to Flash Talk Session

Abstract ID: 1560

Discovery of Pan-Proteomic Common Putative Drug Targets and Vaccine Candidates against Bacterial Urinary Tract Infections

Nalamolu Ravina Madhulitha, Mulpuru Viswajit and Katari Sudheer Kumar*

Department of Biotechnology, School of Biotechnology and Pharmaceutical Sciences, Vignan's Foundation for science, Technology and Research (Deemed to be University), Vadlamudi-522213, Guntur.
*Correspondence mail: drksk_bi@vignan.ac.in@gmail.com

ABSTRACT

There are several micro-organisms that can cause urinary tract infections (UTIs), but *Escherichia coli*, *Klebsiella pneumoniae*, *Proteus mirabilis*, *Enterococcus faecalis* and *Staphylococcus saprophyticus are most prevalent ones.* High recurrence rates and rising uropathogen antibiotic resistance pose a serious threat to the economic burden of these illnesses. Drug resistance and the lack of suitable common targets against genetically varied strains made using the current medications to combat a challenging assignment. The invasive versions of these five species are a grave worry for the entire world, hence pan-proteome comparison studies of urinary tract infections (UTIs) aim to provide plausible and conserved targets that function as therapeutic targets in nature. Five UTI pathogen proteomes reference strains were obtained from the National Center for Biotechnology Information (NCBI) file transfer protocol (ftp) server for pan-proteome comparison studies by utilizing inhouse linux shell scripts, basic local alignment search tool (BLAST) and R scripts. A total of 397 conserved proteins of identity (> 30%) were searched for non-human homology analysis against reference proteomes and resulted in 198 proteins that were completely unidentical, among these 14 were unidentical with human gut metagenome and essentiality analysis revealed these proteins are crucial for the survival of UTI pathogens. The highly prioritized 14 proteins are involved in diverse metabolic pathways, more interaction networks and based on cellular localization studies they might proposed as essential, novel and potent targets for diagnostic, therapeutic and prophylactic design strategies.

Keywords: Urinary tract infection, pan-proteome, potential targets and therapeutic targets.

Abstract ID: 1579
Building the Human Ensemble Cell Atlas and Learning the Underlying Unified Coordinate System

Xuegong Zhang[1,2]

[1]MOE Key Lab of Bioinformatics and Bioinformatics Division of BNRIST, Department of Automation, Tsinghua University, Beijing, China;
[2]School of Life Sciences and School of Medicine, Tsinghua University, Beijing, China.

Abstract: Refer to Additional Flash Tallks

Abstract ID: 1581
Applications of Algebraic Topology to the Detection of Ventricular Tachycardia

Giacomo Pugliese, William Song, Justin Zhang Under the direction of Dr.Krassimir Penev Bergen County Academies

Abstract
Algebraic topology is a powerful tool that can be applied in a variety of settings. Recently, it has found many applications, including to Oncology, 3D shape segmentation, gravitational wave detection, and the classification of handwritten digits. In this paper, we study the applications of persistent homology and simplicial complexes, specifically the Vietoris Rips Complex, to the modeling of ECG data of patients with Ventricular Tachycardia, a condition in which the ventricles of the heart beat abnormally quickly. We then compare the results to those from healthy controls, and examine the differences between the persistence diagrams and persistence images generated from the homology.

Abstract ID: 1609
Pan-proteome relation for targets screening among critical fungal group pathogens

Katari Sudheer Kumar[1,*], Nalamolu Ravina Madhulitha[1], Mulpuru Viswajit[1]

[1]Department of Biotechnology, School of Biotechnology and Pharmaceutical Sciences, Vignan's Foundation for Science, Technology and Research (Deemed to be University), Vadlamudi-522213, Guntur, India
[*]Corresponding author: drksk_bi@vignan.ac.in

Abstract
Fungi are the group of uni or multi cellular eukaryotic heterotrophs alike as bacteria, these microbes also act as friends and foes to human health. World health organization (WHO) on 25-10-2022 released a list of fungal group pathogens (FGP) that are invasive, lethal and microscopic in nature. The list comprises 19 most common fungal pathogens that fall in 3 priority tiers based on public health and due to emerging drug resistant nature, among them the lethargic are critical fungal group pathogens (cFGP) containing 4 species

namely Cryptococcus neoformans, Aspergillus fumigatus, Candida albicans and Candida auris. Research in these organisms is still in infancy stage because all of these organisms' sequences are still partial and yet to be completed. The invasive forms of these four species are of alarming global concern, hence an intention to propose putative and conserved targets which act as therapeutic targets in nature by pan-proteome comparison studies of cFGP strains. The reference strains of four cFGP proteomes from the National Center for Biotechnology Information (NCBI) file transfer protocol (ftp) server were retrieved for pan-proteome comparison studies by utilizing in-house linux shell scripts, basic local alignment search tool (BLAST) and R scripts. 2,220 conserved proteins of identities > 30% from the proteomes were searched for non-human homology analysis resulted in 238 proteins that were completely unidentical, among these 149 were unidentical with human gut metagenome and essentiality analysis revealed 21 proteins that are crucial for the survival of cFGP. The highly prioritized 21 proteins are involved in various metabolic pathways, and more interaction networks, and based on cellular localization studies they might propose as essential, novel and potent targets for diagnostic, therapeutic, and prophylactic design strategies.

Keywords cFGP, pan-proteome, conserved proteins, diagnostic targets, prophylactic targets and therapeutic targets.

Abstract ID: 1617
Deep Transfer Learning of Cancer Drug Responses by Integrating Bulk and Single-cell RNA-seq data

Junyi Chen[1,*], Xiaoying Wang[2,*], Anjun Ma[1,3,$], Qi-En Wang[4], Bingqiang Liu[2], Lang Li[1], Dong Xu[5], Qin Ma[1,3,$]

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA
[2] Department of Mathematics, Shandong University, Shandong 250100, China
[3] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA.
[4] Department of Radiation Oncology, Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
[5] Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[*] These authors contributed equally
[$] To whom correspondence should be addressed

Abstract: Refer to Additional Flash Talks

Abstract ID: 1625
A data-mining approach to explore protein alterations in Rare Neurological Diseases
Author list

Aurelia Morabito[1,2], Giulia De Simone[1,3], Silvia Schiarea[4], Manuela Ferrario[2], Roberta Pastorelli[1], Laura

Brunelli[1]

[1] Laboratory of Metabolites and Proteins in Translational Research, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, 20156 Milan, Italy
[2] Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy
[3] Department of Biotechnologies and Biosciences, Università degli Studi Milano Bicocca, 20126 Milan, Italy
[4] Laboratory of Environmental Epidemiological Indicators, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, 20156 Milan, Italy

Abstract
Rare neurological diseases (RNDs) are a group of conditions that involve the degeneration of specific neuronal populations, causing distinct clinical features. These diseases cause significant difficulties for patients, their families, and society, but current diagnostic and treatment methods are inadequate for many RNDs patients. To improve the development of effective treatments, it is essential to gain a deeper understanding of the underlying mechanisms of these disorders. Research has shown that similar molecular pathways may be involved in different RNDs, and that mutations in the same gene can lead to different clinical conditions. Thus, relying solely on clinical symptoms and genetic information is not enough to fully understand these disorders. A label-free mass spectrometry (MS)-based proteomics analysis was conducted on peripheral blood mononuclear cells (PBMCs) from 50 cerebellar Multiple System Atrophy (MSA-C) patients, 49 Spinocerebellar Ataxia (SCA2) patients, and 50 control subjects. MS data were analyzed by MaxQuant software and MS/MS spectra searched against the Human UniProt FASTA database by the Andromeda search engine. A total of 1613 proteins were identified, 276 of which being detected in more than half of the samples and therefore kept for subsequent analyses. Data was 0-imputed and standardized to account for differences in scale. Then, Linear Discriminant Analysis (LDA) was carried out to identify the proteins driving the discrimination between groups, with the aim of finding the major protein alterations responsible for phenotypic differences. Features were ranked based on the absolute value of their coefficient in the three one-vs-all models. Proteins with a high coefficient in all the models were investigated in the biological RNDs context through MetaCore, a software suite for network analysis. Proteins involved in signal transduction and immune functions were highlighted. Moreover, machine learning models have been employed to validate these proteins' discriminative performance.

Keywords: Proteomics, Mass Spectrometry, Rare Neurological Diseases, Data mining, Machine learning

---

Abstract ID: 1628
A Multi-faceted Mining Tool for Knowledge and Data Discovery for Cancer Research

Avisha Das[1], Omer Anjum[1], Chiamaka Diala[1], W. Jim Zheng[1]

[1] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.

Abstract

Over the decades, extensive research in the field of cancer studies has helped us accumulate knowledge on the biological processes involved in cancer onset, growth, and spread in the body. Such insights into the disease have in turn resulted in effective and targeted treatments and strategies to prevent and contain the disease. The development of safe and effective methods and techniques for diagnosing, treating, preventing, and curing cancer is a crucial aspect of cancer study and research. Additionally, it is equally important to ensure proper education and comprehensive dissemination of knowledge and resources to address the disease effectively. To ensure this, we propose an end-to-end, fully automated data discovery tool for knowledge extraction on cancer research. Recent advances in natural language processing have greatly improved academic literature search platforms (Google Scholar, PubMed).[1] However these systems are limited, performing mostly keyword-based retrieval – a quick search on PubMed.gov with the keyword "cancer" returns approximately 3.5M related full text articles published between 1951 to 2022.[2] Manual verification of relevancy between such a long list of retrieved articles and user query is time consuming and tedious. This creates a need for a systematic and automated pipeline that can retrieve relevant, impactful, and important knowledge from the vast volume of resources. We propose to utilize multiple facets of academic research paper and related meta-data – making use of literature content, citation, and co-authorship networks, to suggest related knowledge. Using such multi-faceted knowledge networks derived from biomedical literature meta-data and content, we design a pipeline for knowledge and data discovery that leverage transformer based large language models like GPT and BERT and adapt it for targeted content mining through supervised fine-tuning. The proposed pipeline results in a much distilled and relevant collection of resources for a given user query. The main contributions of this work are: 1) Through empirical evaluations, we determine the effectiveness and scope of existing multi-faceted approaches in retrieving relevant knowledge resources through content- and network-based methods; 2) Using an experimental task-based evaluation trial, we measure the effectiveness and correctness of the pipeline in mining cancer research literature; and 3) Finally, we build an automated online tool for fast retrieval and ranking of discovered and curated knowledge for widespread use with a continued feedback mechanism for continual improvement.

Keywords: Deep Learning, Textual Mining, Biomedical Literature Mining, Cancer Research

---

Abstract ID: 1655
Supervision versus in-context learning for Mobility functional status information

Tuan Dung Le[1,3], Zhuqi Miao[2], Thanh Thieu[3]

[1]Department of Computer Science and Engineering, University of South Florida, Florida, FL, USA; [2]School of Business, SUNY at New Paltz, New York, NY, USA; [3]Department of Machine Learning, Moffitt Cancer Center and Research Institute, Florida, FL, USA

Abstract
Limitation in physical function reserve, or frailty, is an important factor in cancer treatment and palliative care of elderly patients 65 years or older. Current practice evaluates function using either suboptimal performance status scales ECOG and Karnofsky, or sporadic geriatric measures ADL and IADL. There is no method that uses Natural Language Processing to recognize description of function in clinical notes and prospectively follow them for changes. Using the National NLP Clinical Challenges (n2c2) research dataset,

we begin by constructing a pool of candidate sentences through keyword expansion on all n2c2 sentences. Next, we use pool-based query-by-committee sampling weighted by density representativeness active learning method to select the most informative sentences for human annotation. Each week, two medical personnel annotate sentences on four entity types: Mobility, Action, Assistance, and Quantification. The labeled data is then used to train Bidirectional Encoder Representations from Transformers (BERT) and Conditional Random Field (CRF) models. Predictions from these models on the remaining candidate sentences serve as the basis to select a new batch of sentences for the next annotation iteration. After repeating the active learning cycle for 8 months, we have obtained a dataset of 3,670 sentences that includes a total of 11,102 entities, distributed as follows: 5,520 Action entities, 5,032 Mobility entities, 300 Assistance entities, and 550 Quantification entities. We first evaluate the dataset by training a BERT-based name-entity recognition model using a 5-fold cross-validation setting. The obtained harmonic F1-scores are as follows: 0.81 for Action, 0.66 for Mobility, 0.62 for Assistance, and 0.66 for Quantification. We observed that the model has the highest F-1 score for Action entity, indicating that it is the easiest entity for the model to recognize. However, the model's performance is relatively poor for Assistance and Quantification entities due to their sparsity in the dataset. In addition, we also evaluate the performance of entity recognition in a low-resource setting by applying in-context learning technique with ChatGPT and GPT4 models. We evaluate on smaller curated dataset due to the cost of OpenAI API call. This method shows promising results by providing the large language models with a limited number of examples. The performance of these model still lags behind that of supervision models. Results show the potential applicability of supervision entity recognition and large language models in extracting mobility functioning information from clinical text. We plan to expand investigation to the other function domains to more thoroughly support geriatric assessment.

Keywords: functional status, mobility, electronic health records, clinical notes, n2c2 research datasets natural language processing

---

Abstract ID: 1662
HiC4D: Forecasting spatiotemporal Hi-C data with residual ConvLSTM

Tong Liu[1], Zheng Wang[1]

[1]Department of Computer Science, University of Miami, Coral Gables, Florida, USA

Abstract: Refer to Additional Flash Talks

Abstract ID: 1705
Enhanced Gene Interaction Analysis and Pathway Reconstruction through Iterative Prompt Refinement by ChatGPT

Yibo Chen1,2, Mihail Popescu1, 3, Dong Xu1,2,4

[1] Institute for Data Science and Informatics;

[2] Bond Life Sciences Center;
[3] School of Medicine;
[4] Department of Electrical Engineering and Computer Science; University of Missouri, Columbia, MO, USA.

Abstract: Refer to Flash Talk Session

Abstract ID: 1712

Utilizing Retrospective Administrative Health Claims Data to Investigate Montelukast as a Repurposable Drug Candidate for Multiple Sclerosis Treatment

Astrid M Manuel[1], Assaf Gottlieb[1], Leorah Freeman[2], Zhongming Zhao[1,3]

[1] Center for Precision Health, McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, TX; [2] Neurology Department, Dell Medical School, University of Texas at Austin, TX; [3] Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, TX

Multiple sclerosis (MS) is a complex autoimmune disease and neurological condition characterized by immune-mediated demyelination of the central nervous system. MS presents with brain and spinal cord lesions, which lead to an array of neurological symptoms that typically arise during periods of relapse. According to administrative health claims data, nearly one million people suffer from MS in the United States alone. Despite recent advancements in MS therapeutics, many patients continue to experience breakthrough disease. Previously, we have linked genetic risk factors of MS to existing drug targets and identified repurposable drug candidates for MS. Here, we evaluate a particular drug candidate of interest that stemmed from our previous work, montelukast, a leukotriene receptor antagonist currently indicated for asthma and allergic rhinitis. We designed a retrospective case-control study utilizing real-world clinical data from Optum's de-identified Clinformatics® DataMart (CDM) and IQVIA PharMetrics® Plus for Academics to assess the outcomes of montelukast in MS patients. Inclusion criteria for the cohorts of interest included MS patients, classified with a validated phenotyping algorithm, between ages 18 and 65, and with at least two years of claims data for the longitudinal follow-up analysis. Our main outcome measure was the number of MS relapses, characterized by inpatient hospitalization claims and corticosteroid prescription claims within the two-year period. We identified 691cases in CDM dataset and 301 cases in PharMetrics Plus for Academics, with montelukast prescription claims for a two-year period and medication adherence (proportion of days covered > 0.8). The controls group of this study included MS patients without any montelukast prescriptions and following inclusion criteria, which included 51,237 controls in CDM dataset and 16,647 controls in PharMetrics Plus for Academics dataset. Due to the imbalance between cases and controls, we emulated several randomized clinical trials (RCT), considering the same treatment group and subsets of randomly sampled control patients. For each emulated RCT, we used a doubly robust estimator to adjust for censored patients and multiple confounders, including demographics (age, sex, race), and comorbidities. We observed statistically significant (FDR < 0.05) reduction of relapses in MS patients with montelukast treatment in 95% (224/236) of emulated trials. Importantly, there was an overall 27.8% average reduction in relapses for patients on montelukast. Our

findings provide substantial real-world evidence to drive a drug repurposing strategy for MS, which warrants the need for further investigation.

Keywords: Multiple sclerosis (MS), drug repurposing, montelukast, administrative health claims data, doubly robust estimator, clinical informatics

---

Abstract ID: 1714

Computational Modeling of the Biphasic Depletion of Ovarian Follicle Reserve and Environmental Effects on Ovarian Aging

Sarahna A. Moyd[1], Audrey J. Gaskins[2], Shuo Xiao[3], and Qiang Zhang[1]

[1]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University Atlanta, Georgia, USA;
[2]Department of Epidemiology, Rollins School of Public Health, Emory University Atlanta, Georgia, USA;
[3]Department of Pharmacology and Toxicology, Ernest Mario School of Pharmacy, Environmental and Occupational Health Sciences Institute, Rutgers University, Piscataway, New Jersey, USA.

Abstract

Human ovaries begin development in utero. Through oogenesis, the numbers of oocytes and primordial follicles peak to a few million during fetal development, then decline to hundreds of thousands per ovary at birth. These primordial follicles do not regenerate and are thus regarded as the ovarian reserve. Over the life course, the reserve continues to deplete, due to atresia and activation, until menopause when about 1000 primordial follicles remain. Exposure to chemotherapy drugs and environmental pollutants can accelerate follicular depletion potentially leading to a greater risk of early menopause, primary ovarian insufficiency (POI), and infertility. Physiologically, the ovarian reserve is depleted in a biphasic pattern – a slow decrease from birth to mid-30s, followed by a fast decrease to menopause. While this depletion pattern has been described with empirical mathematical formulations, rarely is it modeled mechanistically. A mechanic model that can characterize the dynamics of follicular depletion throughout the life course will help researchers better understand and predict the impact of chemical exposures on ovarian aging. Here we propose a minimal mechanistic model, which includes (1) a zero-order feedforward inhibition of primordial follicle activation by a local autocrine/paracrine inhibitory factor secreted by the primordial follicles, and (2) a high-gain feedback inhibition of primordial follicle activation by the anti-Müllerian hormone (AMH) secreted by the growing (primary, secondary, and early antral) follicles. The model is configured such that the two regulatory processes prevent primordial follicles from premature overactivation in early and late reproductive life stages, respectively. The model recapitulates the biphasic depletion curve and predicts a constant supply of growing follicles through most of the reproductive active life stage. The model further predicts that the size of the initial primordial follicle pool plays a significant role in determining the menopause age, while unilateral ovariectomy has a minimal effect. Simulations of transient perturbation by chemotherapy drugs, which can promote the atresia of primordial and/or growing follicles, suggest that exposure at earlier ages will have a larger impact on ovarian reserve and menopausal timing than exposure at later ages. Simulations of chronic environmental chemical exposures suggest that chemicals directly promoting primordial follicle atresia are more damaging than chemicals directly promoting growing follicle atresia or inhibiting AMH in advancing menopause age. Future elaborations of the computational model

with integration of in vitro toxicity testing data may help predict the impact of reproductive toxicants on ovarian aging.

Keywords systems biology, biological simulation, modeling, health informatics

---

Abstract ID: 1732
Inferring Single-Cell 3D Chromosomal Structures Based on the Lennard-Jones Potential

Mengsheng Zha[1], Nan Wang[2], Chaoyang Zhang[3], Lluis Morey[1], Zheng Wang[4]

[1] Sylvester Comprehensive Cancer Center, University of Miami, 1475 NW 12th Ave, Miami, FL 33136
[2] Department of Computer Science, New Jersey City University, 2039 Kennedy Blvd, Jersey City, NJ 07305, USA
[3] School of Computing Sciences and Computer Engineering, University of Southern Mississippi, 118 College Dr, Hattiesburg, MS 39406, USA
[4] Department of Computer Science, University of Miami, 1320 S Dixie Hwy,
Coral Gables, FL, USA.

Abstract
Reconstructing three-dimensional (3D) chromosomal structures based on single-cell Hi-C data is a challenging scientific problem due to the extreme sparseness of the single-cell Hi-C data. In this research, we used the Lennard-Jones potential to reconstruct both 500 kb and high-resolution 50 kb chromosomal structures based on single-cell Hi-C data. A chromosome was represented by a string of 500 kb or 50 kb DNA beads and put into a 3D cubic lattice for simulations. A 2D Gaussian function was used to impute the sparse single-cell Hi-C contact matrices. We designed a novel loss function based on the Lennard-Jones potential, in which the ε value, i.e., the well depth, was used to indicate how stable the binding of every pair of beads is. For the bead pairs that have single-cell Hi-C contacts and their neighboring bead pairs, the loss function assigns them stronger binding stability. The Metropolis–Hastings algorithm was used to try different locations for the DNA beads, and simulated annealing was used to optimize the loss function. We proved the correctness and validness of the reconstructed 3D structures by evaluating the models according to multiple criteria and comparing the models with 3D-FISH data.

Keywords: 3D genome; single-cell Hi-C; 3D chromosomal structure; Lennard-Jones potential

---

Abstract ID: 1743
Integrated Spatial Multi-omics Analysis Based on MALDI Data

Xin Ma[1,3], Cameron Shedlock[2,3], Harrison Clarke[2,3], Roberto Ribas[2,3], Terrymar Medina[2,3], Tara R. Hawkinson[2,3], Shannon Keohane[2,3], Craig W. Vander Kooi[2,3], Matthew S. Gentry[2,3], Li Chen[1,3], Ramon Sun[2,3]

[1]Department of Biostatistics, University of Florida, Gainesville, FL, USA;

[2]Department of Biochemistry and Molecular Biology, College of Medicine, University of Florida, Gainesville, FL, USA;
[3]Center for Advanced Spatial Biomolecule Research, University of Florida, Gainesville, FL, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1770
A multimodal neuroimaging-based risk score for Alzheimer's disease by combining clinical and large N>37000 population data

Elaheh Zendehrouh[1,2], Mohammad SE. Sendi[2,3], Vince D. Calhoun[1,2]

[1] Department of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA, USA;
[2] Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA;
[3] McLean Hospital and Harvard Medical School, Boston, MA, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1774
Virtual screening of cyclooxygenase inhibitors from Tinospora cordifolia using the machine learning tool

Abraham Peele Karlapudi*

Department of Biotechnology, Vignan's Foundation for Science, Technology and Research, Vadlamudi-522213, Andhra Pradesh, India.
* Corresponding author: karlapudiabraham@gmail.com

Abstract
Tinospora cordifolia have a variety of compounds and some of these compounds may have anti inflammatory and antioxidant properties. In the present study, we identified the compounds in leaf extract of Tinospora cordifolia through Gas Chromatography-Mass Spectrometry (GC-MS) analysis and found the various metabolites. The compounds are screened virtually using a machine learning model, followed by molecular docking and simulation study for the identification of top hit compounds as inhibitors of cyclooxygenase (COX). The molecular docking revealed that the compound 7,9-Di-tert-butyl-1-oxaspiro (4,5) deca-6,9-diene-2,8-dione (CID:545303) exhibited the lowest binding energies of −7.1 and -6.8 kcal/mol against COX 1 and COX 2 respectively. The interactions are favored by hydrogen bonding and hydrophobic interaction inside the binding pocket. The 100ns MD simulation study for these compounds was performed to know the stability and found the RMSD around 2Å and, around 1.0 Å with minimal fluctuations indicating a stable complex throughout the simulation of 100 ns. Based on these findings, we proposed 7,9-Di-tert-butyl-1-oxaspiro (4,5) deca-6,9-diene-2,8-dione could be used as a dual inhibitor of

COX enzymes and a drug-like a molecule for treating inflammation after evaluation of their biological properties. The web-based platform developed using Streamlit for large scale prediction of cyclooxygenase inhibitors was deployed using the Heroku cloud application platform.

Keywords: 7, 9-Di-tert-butyl-1-oxaspiro (4,5) deca-6,9-diene-2,8-dione, cyclooxygenase inhibitors, Tinospora cordifolia, Virtual screening, Simulation

---

Abstract ID: 1784
FLUXestimator: a webserver for predicting metabolic flux and variations using transcriptomics data.

Alex Lu[1,2], Zixuan Zhang[2+], Haiqi Zhu[2,3+], Pengtao Dang[2,4+], Jia Wang[2,3], Wennan Chang[2], Xiao Wang[2,3], Norah Alghamdi[2], Yong Zang[2,5], Wenzhuo Wu[6], Yijie Wang[3], Yu Zhang[2]*, Sha Cao[2,5]*, Chi Zhang[2]*

[1] Park Tudor School, Indianapolis, IN, US 46240
[2] Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, US 46202
[3] Department of Computer Sciences, Indiana University, Bloomington, IN, US 47405
[4] Department of Electric Computer Engineering, Purdue University, Indianapolis, IN, US 46202
[5] Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, US 46202
[6] Department of Industrial Engineering, Purdue University, West Lafayette, IN, US 47907

Abstract: Refer to Session "Artificial Intelligence on Big Data: Promise for Early-stage Trainees".

Abstract ID: 1803
Integrating Hydrogen Bonding Information into Graph Neural Networks for Protein Structure Classification

Yi-Shan Lan[1], Tsung-Yi Ho[2]

[1]Department of Computer Science, Institute of Information Systems & Applications, National Tsing Hua University, Hsinchu 30013, Taiwan
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong

Abstract: Refer to Additional Flash Talks

Abstract ID: 1809
High-resolution Refinement of Population Affinity Estimation through Deep Learning Technology with Craniometric Data

Jinyong Pang[1], Xiaoming Liu[1]

[1]USF Genomics & College of Public Health, University of South Florida, Tampa, FL, USA.

Abstract

Making valid inference on the social identity for the unidentified biological remains of human is essential in forensic investigations. The cranium, as informative skeleton remains, plays a vital role in population affinity estimation based on cranial characteristics. For craniometric population affinity estimation, traditional statistical methods and machine learning models, including geometric morphometrics (GMM), linear discriminant analysis (LDA) and tree-base model (TM), are commonly used. The ongoing debate regarding the use of ancestry estimation in forensic anthropology has emphasized the need for techniques that are effective in identifying diverse population groups with greater precision. Deep leaning (DL) models as the novel approach, more capable in handling complicated classification tasks than traditional machine learning methods in many fields, opens up new vistas of this practice for forensic anthropologists. In our research, several deep feedforward neural network models for population affinity estimation were developed by using Howells' craniometric data set, in which 82 craniometric measurements from 2,412 human crania from 26 populations were covered. We obtained average accuracy of 96% for 6-group (Europe, Africa, Austro-Melanesia, Polynesia-Micronesia, America, and East Asia) affinity estimation, 95% for 11-group (East Africa, West Africa, South Africa, Greenland, Siberia, East-Southeast Asia, Andaman, America, Polynesia-Micronesia, Australia-Melanesia, Central-North Europe) affinity estimation, and 93% for 26-population affinity estimation. Additionally, the impact of each feature on the predictive model are quantified based on Shaply values, which take into account the underlying interactions and dependencies between features, allowing for a more nuanced understanding of the DL model's behavior in population affinity estimation. The results demonstrate that deep learning technology can help achieve reasonable to high accuracy for fine-grained population affinity estimation and highlight its potential to significantly advance the understanding of craniometric data.

Keywords: ancestry estimation, craniometrics, deep learning

Abstract ID: 1810

Pan-Cancer Analysis for Subgroup Discovery and Drug Repositioning

Zainab Al-Taie[1,8,9], Jonathan Mitchem[1,4,7], Mark Hannink[5,6], Jussuf T. Kaifi[4,7], Christos Papageorgiou[4], Chi-Ren Shyu[1,2,3,*]

[1]Institute for Data Science & Informatics, University of Missouri, Columbia, MO 65211, USA
[2]Electrical Engineering and Computer Science Department, University of Missouri, Columbia, MO 65211, USA
[3]Department of Medicine, School of Medicine, University of Missouri, Columbia, MO 65212, USA
[4]Department of Surgery, School of Medicine, University of Missouri, Columbia, MO 65212, USA
[5]Department of Animal Sciences, Bond Life Sciences Center, University of Missouri, 1201 Rollins Street, Columbia, MO, 65211
[6]Department of Biochemistry, University of Missouri, Columbia, Missouri, USA
[7]Harry S. Truman Memorial Veterans' Hospital, Columbia, MO 65201, USA
[8]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA 9Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Abstract
Cancer is a group of diseases. Each cancer type has a different mortality rate, and patients of each type have heterogeneous responses to the treatment. However, all these types of cancer are characterized by uncontrolled growth and the spread of abnormal cells. This suggests the existence of common mechanisms among these types in addition to the unique mechanisms for each type. In this study, we are implementing our drug repositioning and subgroup discovery method to find homogeneous subgroups in cancer across different cancer types and reposition drugs based on the genotypic features of each subgroup. A pan-cancer analysis was conducted to find common mechanisms. From The Cancer Genome Atlas (TCGA), the genotypic and phenotypic data for 3983 patients across cancer types in 11 organs were obtained. The phenotypic data are the clinical variables, and the genotypic data are the RNA-seq data for each type. The data were combined and preprocessed. Then, the subgrouping algorithm was implemented on this dataset to find homogeneous subgroups within the heterogeneous cancer population. After filtering the resulting subgroups, the subgroups with gender as a significant phenotypic variable were selected for further analysis. To have a comparable result, the subgroups that appeared in both genders were chosen to be analyzed. This resulted in 50 subgroups. The drugs were recommended for the top subgroups based on the uniquely abnormal genes in each subgroup. The analysis of the recommended drugs for each subgroup across different types of cancer showed that these drugs could be repurposed for more than one cancer type. We found that the majority of these drugs are FDA-approved drugs for cancer and our analysis showed the existence of shared mechanisms that these drugs can target in additional types of cancer. Further analysis is needed in the context of wet-lab experiments to validate these results before recommending them for clinical use.

Abstract ID: 1816
Decentralization of Brain age Estimation with Structural Magnetic Resonance Imaging Data

Sunitha Basodi[1], Rajikha Raja[2], Bhaskar Ray[1,3], Harshvardhan Gazula[4], Jingyu Liu[1,3], Eric Verner[1] and Vince D. Calhoun[1,3,5]

[1] Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA
[2] St. Jude Children's Research Hospital, Memphis, TN, USA
[3] Department of Computer Science, Georgia State University, Atlanta, GA, USA
[4] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[5] Department of Psychology, Georgia State University, Atlanta, GA, USA

Abstract: Refer to Additional Flash Talks

Abstract ID: 1838
PROMPT BIOINFO. CASE STUDY: Shotgun Metagenomic Data Analysis

Zhu Xing[1,2], Qiyun Zhu[1,2*]

[1] School of Life Sciences, Arizona State University, Tempe, Arizona 85281, USA
[2] Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, Arizona 85281, USA
[*] Contact author: qiyun.zhu@asu.edu

Abstract: Refer to Additional Flash Talks

Abstract ID: 1839
A Novel Statistical method for Differential Analysis of Single-Cell Chromatin Accessibility Sequencing Data

Fengdi Zhao[1], Xin Ma[1], Li Chen[1]
[1]Department of Biostatistics, College of Public Health and Health Professions University of Florida, Gainesville, FL, USA

Abstract
Single-cell ATAC-seq sequencing data (scATAC-seq) has been a widely adopted technology to investigate the chromatin accessibility on the single cell level. Analyzing scATAC-seq can provide valuable insights in identifying cell populations and revealing the epigenetic heterogeneity across cell populations in different biological context. One important aspect of scATAC-seq data analysis is performing differential chromatin accessibility (DA) analysis, which will help identify cell populations and reveal epigenetic heterogeneity. While numerous differential expression methods have been proposed for single-cell RNA sequencing data, DA methods for scATAC-seq data are under-developed and remains a major challenge due to high sparsity and high dimensionality of the data. To fill the gap, we introduce a novel and robust a zero-inflated negative binomial framework named scAC-DA for DA analysis. The model links the prevalence, dispersion, and mean parameter to covariates such as cell populations, treatment conditions and batch effect. The statistical inference is based on EM algorithm and the dispersion parameter is shrunk using an empirical Bayes method to stabilize the parameter estimation by leveraging information from other accessible chromatin regions in the genome. Consequently, we performed both simulation studies and real data applications, which demonstrate the superiority of scAC-DA compared to existing approaches.

Keywords: Chromatin accessibility, single cell sequencing, empirical bayes

---

Abstract ID: 1868
Multimodal machine learning combining image and textual data to predict rare genetic disorders

Da Wu[1], Jingye Yang[1], Kai Wang[1]

[1]Children's Hospital of Philadelphia, Philadelphia, PA, USA.

Abstract: Refer to Flash Talk Session

Abstract ID: 1870

Single-cell Mayo Map (scMayoMap): an easy-to-use tool for cell type annotation in single-cell RNA-sequencing data analysis

Lu Yang[1,2,†], Yan Er Ng[3,†], Haipeng Sun[4], Ying Li[5], Lucas C.S. Chini[3], Nathan K. LeBrasseur[3,6,*], Jun Chen[1,2,*], Xu Zhang*

[1]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, 55905, USA.
[2]Center for Individualized Medicine, Mayo Clinic, Rochester, MN, 55905, USA.
[3]Robert and Arlene Kogod Center on Aging, Mayo Clinic, Rochester, MN, 55905, USA.
[4]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ, 08901, USA.
[5]Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, FL, 32224, USA.
[6]Department of Physical Medicine and Rehabilitation, Mayo Clinic, Rochester, MN, 55905, USA.
[7]Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, MN, 55905, USA.
[†]These authors equally contributed to this work.
[*]To whom correspondence should be addressed.

Abstract

Single-cell RNA-sequencing (scRNA-seq) has become a widely used tool for both basic and translational biomedical research. In scRNA-seq data analysis, cell type annotation is an essential but challenging step. In the past few years, several annotation tools have been developed. These methods require either labeled training/reference datasets, which are not always available, or a list of predefined cell subset markers, which are subject to biases. Thus, a user-friendly and precise annotation tool is still critically needed. We curated a comprehensive cell marker database named scMayoMapDatabase and developed a companion R package scMayoMap, an easy-to-use single cell annotation tool, to provide fast and accurate cell type annotation. The effectiveness of scMayoMap was demonstrated in 48 independent scRNA-seq datasets across different platforms and tissues. scMayoMap performs better than the currently available annotation tools on all the datasets tested. Additionally, the scMayoMapDatabase can be integrated with other tools and further improve their performance. scMayoMap and scMayoMapDatabase will help investigators to define the cell types in their scRNA-seq data in a streamlined and user-friendly way.

Keywords: scRNA-seq; scMayoMap; scMayoMapDatabase; cell type annotation

---

Abstract ID: 1890

Cell type-specific differential landscape in the human placenta for preeclampsia identified by novel scRNA-seq pipeline

Yuheng Du[1], Qianhui Huang[1], Paula A. Benny[2], Youqi Yang[3], Ryan J. Schlueter[4], Cameron Lassiter[2], Lana X. Garmire[1]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA; [2]University of Hawaii Cancer Center, Epidemiology, Honolulu, HI; [3]Department of Biostatistics,

University of Michigan, Ann Arbor, MI, USA; [4]Department of Obstetrics and Gynaecology, University of Hawaii, Honolulu, HI.

Abstract
Placenta is the unique organ of maternal vs. fetal cells as well as blood vs. tissue cells. Deciphering the cell origins in the placenta presents a unique challenge for single-cell RNA-seq (scRNA-Seq) analysis among all types of human tissues. To address this challenge, we proposed a multi-step scRNA-Seq analytical pipeline to human placental, allowing the identification of the origins of single cells into four categories: maternal vs. fetal, and blood vs. solid tissues. We applied this workflow to five preeclampsia (PE) samples vs. five healthy control samples. We focus on the downstream functional analysis of the solid fetal tissue portion. Bioinformatics analysis shows that PE patients present suppressed placental innate immune functions in a variety of cell types. Moreover, cell-cell communication analysis demonstrates increased connections with Fibroblasts, suggesting fibrosis in the preeclamptic placenta. Gestational age-adjusted pseudo-time differentiation trajectory analysis shows that syncytiotrophoblasts (STB) and extravillous trophoblasts (EVT) in the PE samples are significantly dysregulated, through fate-decision genes, including FN1 and HLA-G for EVT, and CGA, CSH1, and PSG genes. Further, geolocation differences are apparent across the human placenta, varying from the fetal layer, intermediate layer, to the maternal layer, demanding the need for refined transcriptomics analysis within heterogeneous tissue like the placenta. In conclusion, our work sets up the foundation for a more rigorous scRNA-seq analytical framework to dissect the unique challenges of heterogeneity in the placenta.

Keywords: Single-cell, cell origin, preeclampsia, placenta, immunology, trajectory analysis

---

Abstract ID: 1895
Mutated processes predict immune checkpoint inhibitor therapy benefit in metastatic melanoma

Andrew Patterson[1], Noam Auslander[2]

[1]Genomics and Computational Biology Graduate Group, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA, 19104, USA
[2]Program in Molecular and Cellular Oncogenesis, The Wistar Institute, Philadelphia, PA, 19104, USA

Abstract: Refer to Additional Flash Talks

Abstract ID: 1896
Identifying potential radiosensitizers using RSI and the Connectivity Map

Steven Eschrich[1], Eric Welsh[2], Kamran Ahmed[3], Javier TorresRoca[3]

[1]Department of Biostatistics & Bioinformatics; [2]Biostatistics & Bioinformatics Shared Resource
[3]Department of Radiation Oncology, Moffitt Cancer Center, Tampa, FL, USA

Abstract

Background: The radiosensitivity index (RSI) is a tumor-based gene expression signature that is predictive of patient response to radiotherapy and has been validated in 22 datasets from 12 disease sites in over 5000 patients treated with radiation. The genomic-adjusted radiation dose (GARD) uses RSI to estimate the effect of radiation on a tumor. Using RSI/GARD, a subset of tumors is resistant to radiation even at higher doses. Therefore, identifying therapeutics that sensitize these tumors to radiation therapy (radiosensitizers) is an important step in providing effective radiotherapy for all patients. **Methods:** We normalized raw (CEL file) Connectivity Map (CMap) gene expression data (Build02) using the robust multiarray analysis (RMA). This CMap data was used to calculate RSI across the cell lines (HL60, MCF7, PC3, SKMEL5 and ssMCF7/MCF7 variant) and conditions (vehicle control (DMSO) and drug treatment (perturbation)). RSI differences (deltaRSI) were computed across the ~1300 drugs tested (perturbation RSI – vehicle RSI). We used a deltaRSI threshold of 0.15 (approximately 2 sd's from the mean deltaRSI) as significant changes in RSI. The deltaRSI from replicate treatments was summarized using the mean.
Results: 228 unique cell line/drug treatment combinations were identified as potential radiosensitizers (deltaRSI < -0.15) from a total of 3587 total combinations. Of the 228 significant combinations, there were 104 HL60, 106 MCF7 hits, and 18 PC3 potential radiosensitizers (deltaRSI greater than 0.15). There were only 14 (7.1%) drug combinations identified as different in both HL60 and MCF7 cell lines suggesting that radiosensitization may be tissue-specific. The mean vehicle control RSI of the cell lines across the entire dataset indicates baseline radioresistance (HL60=0.536, MCF7=0.701, PC3=0.611, SKMEL5=0.607, ssMCF7=0.527). The mean deltaRSI changes per cell line were ~0.2 (HL60, deltaRSI=0.199, MCF7, deltaRSI=0.194, PC3 deltaRSI=0.184) and at least 15 combinations were greater than 0.3 deltaRSI. Finally, since RSI is composed of gene expression from 10 unique genes, we identified RSI genes with the largest mean differences in perturbation vs. vehicle control among the potential radiosensitizers. Interestingly, both JUN (fold-change: 1.73) and IRF1 (fold-change: 1.81) increased in gene expression after drug treatment.
Conclusion: Drug repurposing using the connectivity map gene expression dataset with the RSI molecular gene signature can be used to screen for potential radiosensitizers. This analysis suggests that radiosensitization may be tissue specific. This process can prioritize potential drug/cell line combinations for further testing and identify specific targets (e.g., JUN, IRF1) for radiosensitization.

Keywords: Drug repurposing, radiation sensitivity, connectivity map, gene expression

---

Abstract ID: 1898

Association between ABCG1/TCF7L2 and type 2 diabetes mellitus: An intervention trial based case-control study

Yinxia Su[1#] Xiangtao Liu[1#] Conghui Hui[2] Hua Yao[3]*

[1] School of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, Xinjiang, China; [2] School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang, China; [3] School of Health Management, Xinjiang Medical University, Urumqi, Xinjiang, China.

Abstract

Background and Objective: Type 2 diabetes mellitus (T2DM) is the result of both genetic and environmental factors. Environmental factors may contribute to the occurrence and development of T2DM by influencing epigenetic modification. DNA methylation is a major modification mode and an important

regulatory mechanism of epigenetic inheritance, which is considered to be an important phenotypic outcome and marker of disease progression. In this study, we focused on the methylation sites of TCF7L2 and ABCG1 genes that are most strongly associated with T2DM. By conducting intervention experiments in Uyghur population, which has been less studied, to analyze the potential functions of SNP-CG sites rs7901695 and cg06500161 of the above two genes as biomarkers in the development of T2DM, and provide evidence for personalized health management of T2DM in Uyghur people. **Methods:** 320 patients with T2DM and 332 patients without T2DM were treated with dietary pagoda-based health education intervention. The demographic data and basic physical biochemical indexes before and after intervention were collected by questionnaire survey and physical biochemical examination. SNP typing was performed by Taqman-MGB probe method, and gene methylation was detected by pyrosequencing method. **Results:** 1. The genotypes of SNP sites corresponding to the methylation site cg06500161 of ABCG1 gene were all CC type without gene polymorphism, so the polymorphism of this gene locus was not analyzed. The rs7901695 genotypes of TCF7L2 gene were TT, TC and CC. But only 98 out of 332 samples contained the C allele, and the methylation modification was limited to cytosine in GC sequence. Due to the small sample size, the correlation analysis between methylation level and T2DM at this site was not conducted. 2. The rs7901695 genotype difference of TCF7L2 was statistically significant between the case group and the control group ($P < 0.05$). After adjusting for covariates (smoking, alcohol consumption, exercise, FPG, obesity and hypertension), genotype of rs7901695 in TCF7L2 gene was associated with genetic susceptibility to T2DM in addition (TC vs TT, $P = 0.047$; CC vs TT, $P = 0.010$), dominant ($P = 0.015$) and recessive ($P = 0.039$) models. 3. Before intervention, there were significant differences in the intake of water between the case group and the control group ($P < 0.05$); After intervention, there was statistical significance in the intake of coarse grains, fruits, aquatic products, eggs, dairy products, soy products, nuts, edible oils and water between the case group and the control group (Ps < 0.05). Logistic regression analysis showed that methylation of ABCG1 gene was correlated with T2DM susceptibility after adjustment of covariable before intervention ($P = 0.015$, OR: 1.023; 95%CI: 1.004~1.041) but not after intervention. 4. Generalized Multifactor Dimensionality Reduction (GMDR) showed rs7901695 locus of TCF7L2 gene and cg06500161 locus of ABCG1 gene had interaction with hypertension, dyslipidemia, abdominal obesity and obesity, and also had interaction with drinking, smoking and exercise. **Conclusions:** The polymorphism of rs7901695 site of TCF7L2 gene is associated with the incidence of T2DM in Uyghurs. The interaction of rs7901695 site of TCF7L2 gene and cg06500161 site of ABCG1 gene with environmental factors may increase the risk of T2DM in Uyghurs. The interaction between cg06500161 site of ABCG1 gene and environmental factors on T2DM varied with the intervention. Cg06500161 site of ABCG1 may serve as a biomarker to evaluate the effect of T2DM interventions.

Keywords: ABCG1; TCF7L2; Single nucleotide polymorphism; Methylation; Type 2 diabetes mellitus

---

Abstract ID: 1908

Exploring the impact of structural variants on the genetic etiology of autism spectrum disorder and language impairments

Rohan Alibutud[1,#], Sammy Hansali[1,#], Xiaolong Cao[1,6], Anbo Zhou[1], Vaidhyanathan Mahaganapathy[1], Marco Azaro[1], Christine Gwin[1], Sherri Wilson[1], Steven Buyske[2], Christopher W. Bartlett[3,4], Judy F. Flax[1], Linda M. Brzustowicz[1,5], Jinchuan Xing[1,5,*]

[1] Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[2] Department of Statistics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[3] The Steve Cindy Rasmussen Institute for Genomic Medicine, Battelle Center for Computational Biology, Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, Ohio, USA
[4] Department of Pediatrics, College of Medicine, The Ohio State University, Columbus, Ohio, USA
[5] The Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[6] Current address: Division of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, China

Abstract
Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by restrictive interests and/or repetitive behaviors and deficits in social interaction and communication. ASD has a complex polygenic genetic architecture, and its contributing factors are not yet fully understood, especially large structural variations. In this study we aim to assess the contribution of structural variations, including copy number variants (CNVs), insertions, deletions, duplications, and mobile element insertions, to ASD and related language impairments in the New Jersey Language and Autism Genetics Study (NJLAGS) cohort. We showed a significantly higher number of CNVs in ASD patients (median n = 5) than unaffected individuals (median n = 3, $p < 1e^{-5}$). Within the cohort ~77% of the families contain structural variations that followed expected segregation or de novo patterns and passed our filtering criteria. These structural variations affected 264 brain-expressed genes and can potentially contribute to the genetic etiology of the disorders. Gene Ontology and protein-protein interaction network analysis suggested several clusters of genes in different function categories, such as components of axon and histone modification machinery. Genes and biological processes identified in this study contribute to the understanding of ASD and related neurodevelopment disorders.

Keywords (up to six words): autism, copy number variation, structure variation

---

Abstract ID: 1915
A massive proteogenomic screen identifies thousands of novel peptides from the human "dark" proteome

Xiaolong Cao[1,2], Siqi Sun[2], Jinchuan Xing[2,*]

[1] Division of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong 510280, China;
[2] Department of Genetics, Human Genetic Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

Abstract: Refer to Additional Flash Talks

Abstract ID: 1943

Unified somatic calling and machine learning-based classification enhance the discovery of clonal hematopoiesis of indeterminate potential

Shulan Tian[1,4], Garrett Jenkinson[1,4], Alejandro Ferrer[2], Huihuang Yan[1,4], Joel Morales-Rosado[1,4], Kevin L. Wang[1,4], Terra Lasho[2], Saurabh Baheti[1], Janet E. Olson[5], Linda B. Baughn[6], Mrinal S. Patnaik[2], Wei Ding[2], Konstantinos N. Lazaridis[3,4], and Eric W. Klee[1,4]

[1]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA
[2]Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA
[3]Division of Gastroenterology & Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA
[4]Center for Individuated Medicine, Mayo Clinic, Rochester, MN 55905, USA
[5]Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA
[6]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1986
Genomic disparities between cancers in adolescent and young adults and in older adults

Xiaojing Wang[1,2], Anne-Marie Langevin[3], Peter Houghton[1,4], Siyuan Zheng[1,2]

[1]Greehey Children's Cancer Research Institute, UT Health San Antonio, TX, USA;
[2]Department of Population Health Sciences, UT Health San Antonio, TX, USA;
[3]Department of Pediatrics, UT Health San Antonio, TX, USA;
[4]Department of Molecular Medicine, UT Health San Antonio, TX, USA.

Abstract: Refer to Additional Flash Talks

Abstract ID: 1987
In silico Improvement of Highly Protective Antimalarial Antibodies

Mateo Reveiz[1], Andrew Schaub[1], Young Do Kwon[1], Prabhanshu Tripathi[1], Azza Idris[1,2], Amarendra Pegu[1], Lais Da Silva Pereira[1], Patience Kiyuka[1], Myungjin Lee[1], Tracy Liu[1], Chen-Hsiang Shen[1], Baoshan Zhang[1], Yongping Yang[1], Peter D. Kwong[1], Reda Rawi[1]
[1]Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA. [2] The Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA, 02139, USA

Abstract: Refer to Flash Talk Session

Abstract ID: 1988
Direct Evidence of Metabolic Interactions between PBDEs and Gut Microbes: an In Vitro Metabolomics Study

Yan Jin[1], Kyle Kim[2], Julie De Laloubie[1,3], Kris Weigel[2], Julia Yue Cui[2], Haiwei Gu[1]

[1] Center for Translational Science, Florida International University, Port St. Lucie, FL
[2] Department of Environmental and Occupational Health Sciences, University of Washington, WA
[3] School of Engineering Sciences, Polytech Clermont-Ferrand, Clermont Auvergne University, France

Abstract
Polybrominated diphenyl ethers (PBDEs) were extensively used as flame retardants in various factory products and are still persistent in the environment after they were banned to use. Recently, we showed in various in vivo studies that gut microbiome dysbiosis is involved in PBDEs-induced toxicity; however, it is unknown whether PBDE exposure directly impacts the metabolism of the gut microbiome. In this in vitro study, we used mass spectrometry-based metabolomics approaches to investigate the metabolic interactions between 2 common PBDE congeners (2,2′,4,4′-tetrabromodiphenyl ether [BDE-47], 2,2',4,4',5-pentabromodiphenyl ether [BDE-99]) and selected gut microbe species (Akkermansia muciniphila [AKK] and Clostridium scindens [CS]) that are known to be involved in metabolic diseases, as well as Escherichia coli. (EC), an established control microbe. All microbes were individually cultured in an anaerobic workstation. These bacterial strains were exposed with BDE-47 and BDE-99 of different concentrations (0, 10 μM, or 100 μM). Pathway-specific targeted LC-MS/MS metabolomics was used to examine ~300 aqueous metabolites from ~30 metabolic pathways of biological significance. Our results suggest that both BDE-47 and BDE-99 significantly altered metabolic profiles in all these 3 bacterial strains. For example, in the principal component analysis (PCA) score plot (**Figure 1**), CS after BDE-47 exposure is clearly separated from controls in a dose-dependent manner. Additionally, metabolomics results revealed that BDE-47 and BDE-99 caused different metabolic responses in these gut microbes. For example, CS had increased lactate and decreased kynurenine after BDE-99 exposure, while BDE-47 induced decreased 2-pyrrolidinone and increased urocanic acid in CS. Furthermore, various altered metabolites were found significant in multiple metabolic pathways, especially in glycolysis, TCA cycle, and tryptophan metabolism. A total of 60 metabolic features were determined to distinguish potentially disturbed metabolite markers of BDE-47 and BDE-99 exposure. In conclusion, our findings provide possible biomarkers of toxic effects induced by BDE-47 and BDE-99 and elicit a deeper understanding of the metabolic mechanisms that could be validated in further in vivo studies.

Abstract ID: 1990
RNA-seq Improves SARS-CoV-2 Gene Annotations

Alan Jiang[1], Allen Zhao[2]

[1]Bridgewater Raritan High School, Bridgewater, New Jersey, NJ, USA; [2] The Peddie school, Hightstown, New Jersey, NJ, USA;

Abstract
Gene annotation is a way to identify elements in genomic sequence and to derive meaning of them in GTF (Gene transfer format) or GFF (General Feature Format) file. Due to the high variability of virus genome, the GTF or GFF file in public database may not be sufficient to capture gene features of the most currently popular virus strains. In this study, we developed an RNA-seq analysis method to capture the new gene features and update them in the GTF file by leveraging the most recent public SARS-Cov-2 RNA-seq data. We used nf-core/rnaseq pipeline in Amazon cloud to analyze the RNA-seq data downloaded from NCBI SRA database, providing it with a reference genome and GTF file from NCBI GenBank. After running the pipeline, we received sequence alignment data in the form of bam files. In addition to this, the pipeline also used StringTie to assemble the alignment data in to GTF files. Next, we used GFFread to convert this GTF file and genome sequence into a transcript fasta file, which was then analyzed using transdecoder. Because the annotation file generated by transdecoder did not provide the coordinates for the genes in the scale of genome, we developed our own python script to map them onto the genome. After running the whole analysis pipeline, we annotated the genes with UTRs and introns in addition to exons in the original GTF. We also discovered the introns associated with splicing events in all three SARS-CoV-2 strains. The result suggested that gene annotations of Virus could be expanded by the newly developed RNA-seq analysis method.

Keywords: Gene Annotation, SARS-CoV-2, RNA-seq, GTF/GFF

CONFERENCE LOCATION

Hilton St Petersburg Bayfront

333 1st St SE, Saint Petersburg, FL 33701

The Hilton St. Petersburg is located in downtown St. Petersburg facing the beautiful Tampa Bay. It is at the heart of the St. Petersburg Waterfront District, with all the iconic attractions in St. Petersburg, such as the Salvador Dali Museum, St. Petersburg Pier, and Vinoy Park. While the Hilton St. Petersburg Bayfront has several restaurants and bars, such as the Dali Restaurant and Bar, Starbucks Coffee Bar, and Tangerine restaurant, for onsite dining, it has access to numerous restaurants, cafes, and retails within walking distance of the hotel. Furthermore, the Downtown Looper trolley has a stop at the hotel, providing links to all St. Petersburg downtown attractions. The Central Avenue trolley is three blocks from the hotel, providing links to one of the best beaches in the US, St. Petersburg Beach.

Parking Information

Hotel Parking
Valet parking: $35.00

Public Parking Options
Platinum Parking City Center Tower (north of the hotel):
Address: 100 Second Avenue S, St. Petersburg, Florida 33701
Parking rate: $15 Daily
Website: https://www.downtownstpetersburgparking.com/City-Center-Tower-Parking

Additional Parking Options

Lot and Garage Information is also available at:
The Philadelphia Parking Authority (philapark.org)
Philadelphia Parking - From $10 - Find, Book & Save 60% on Philly Parking(parkwhiz.com)
Philadelphia Parking - Save Up to 50% | SpotHero

Airport Information

Tampa International Airport (TPA): 22 miles from the Hotel.

Hotel Information

Hilton St. Petersburg Bayfront

Map and Directions

Room Rate: $175 Double or King

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

Beaches and Local attractions

Some information on beaches and local attractions can be found here.

SPECIAL ACKNOWLEDGEMENTS

We are grateful for the numerous helps from the following volunteers:

Andi Liu (Chair)	University of Texas Health Science Center at Houston
Keith L. Sanders	University of Texas Health Science Center at Houston
Astrid M. Manuel	University of Texas Health Science Center at Houston
Ruoying Yuan	Washington University in St. Louis
Di Huang	Washington University in St. Louis
Pramod Bharadwaj Chandrashekar	Arizona State University & University of Wisconsin-Madison
Jinyong Pang	University of South Florida
Chang Li	University of South Florida
Weiru Han	University of South Florida
Kun Bu	University of South Florida
Sumarga Sah Tyagi	University of South Florida
Tam Minh Nguyen	University of South Florida
Huu Dang Pham	University of South Florida

MANY THANKS TO OUR SPONSORS!

About Admera Health

Admera Health is a company with laboratory specializing in genomic and bioinformatic services for customers worldwide, including biotech, biopharma, and pharmaceutical companies, universities, and nonprofit institutions, to advance human life and health.

Our Work and Mission

We collaborated with major cancer centers in the US, top pharma companies, and Ivy League universities on projects related to human, animal, and plant health. Originally a spin-off from Genewiz, we transitioned to sequencing laboratory work in biopharma, leveraging their expertise in science and next-generation sequencing (NGS) since 2015. We maintain certifications such as CLIA, CLEP, and CAP accreditation, ensuring high-quality care standards and guidance to customers, and utilize advanced technology and automated liquid handlers to enhance efficiency and service quality. We securely store sequencing data using Illumina BaseSpace Sequence Hub and internal servers.

Our Team

Our team, composed of PhD-level scientists, analyzes, and annotates vast amounts of data, providing meaningful insights to customers engaged in academic research, pharmaceutical studies, and biotech product research.

Overall, we combine cutting-edge technology, expertise in NGS, and a strong bioinformatics team to offer genomic and bioinformatic services to a diverse range of customers in the healthcare and research sectors. For more information, please visit

https://www.admerahealth.com

bi.admerahealth.com (developing)

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.

At Complete Genomics, our mission is to redefine the genomic landscape with accessible and truly complete solutions. Powered by unrivaled technology, we've created a full spectrum of NGS products—from lab automation to sequencing platforms to data analysis—disrupting the traditional way of developing genomic tools and accelerating scientific breakthroughs for scientists everywhere.

Our comprehensive suite of instruments, ranging from low- to ultra-high-throughput genetic sequencers, has generated over 100 petabytes of data for more that 1,300 users worldwide.

Based in San Jose, California since 2005, Complete Genomics has the ability to provide local support across the US as well as to ensure there is always supply available.

Computational and Structural Biotechnology Journal (***CSBJ***) is an online open-access journal publishing research articles and reviews after full peer review (ISSN 2001-0370). With an Impact Factor ***6.15*** and CiteScore ***7.60***, the popularity of *CSBJ* has been increasing steadily since its launch and been ranked among the top journals in the field. The journal welcomes the submission of manuscripts that meet the general criteria of significance and scientific excellence, and enables the rapid publication of papers under the following sections:

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL — GENERAL

THE SMART HOSPITAL SECTION

***CSBJ*: General Section** places a strong emphasis on functional and mechanistic understanding of how molecular components in a biological process work together through the application of computational methods.

***CSBJ*: The Smart Hospital Section** places a strong emphasis on new digital and automated technologies transforming health and care systems. Clinical and piloting data may provide practical insights on implementing digital and automated technologies in real-world smart hospital settings, but they are not a prerequisite for publication in this section.

For more information, visit the journal homepage: **csbj-elsevier.com**

Patterns is a premium open access journal from Cell Press, publishing ground-breaking original research across the full breadth of data science. Works published at the journal are expected to present significant new advances, and to share data and code in a manner that exemplifies open and FAIR science.

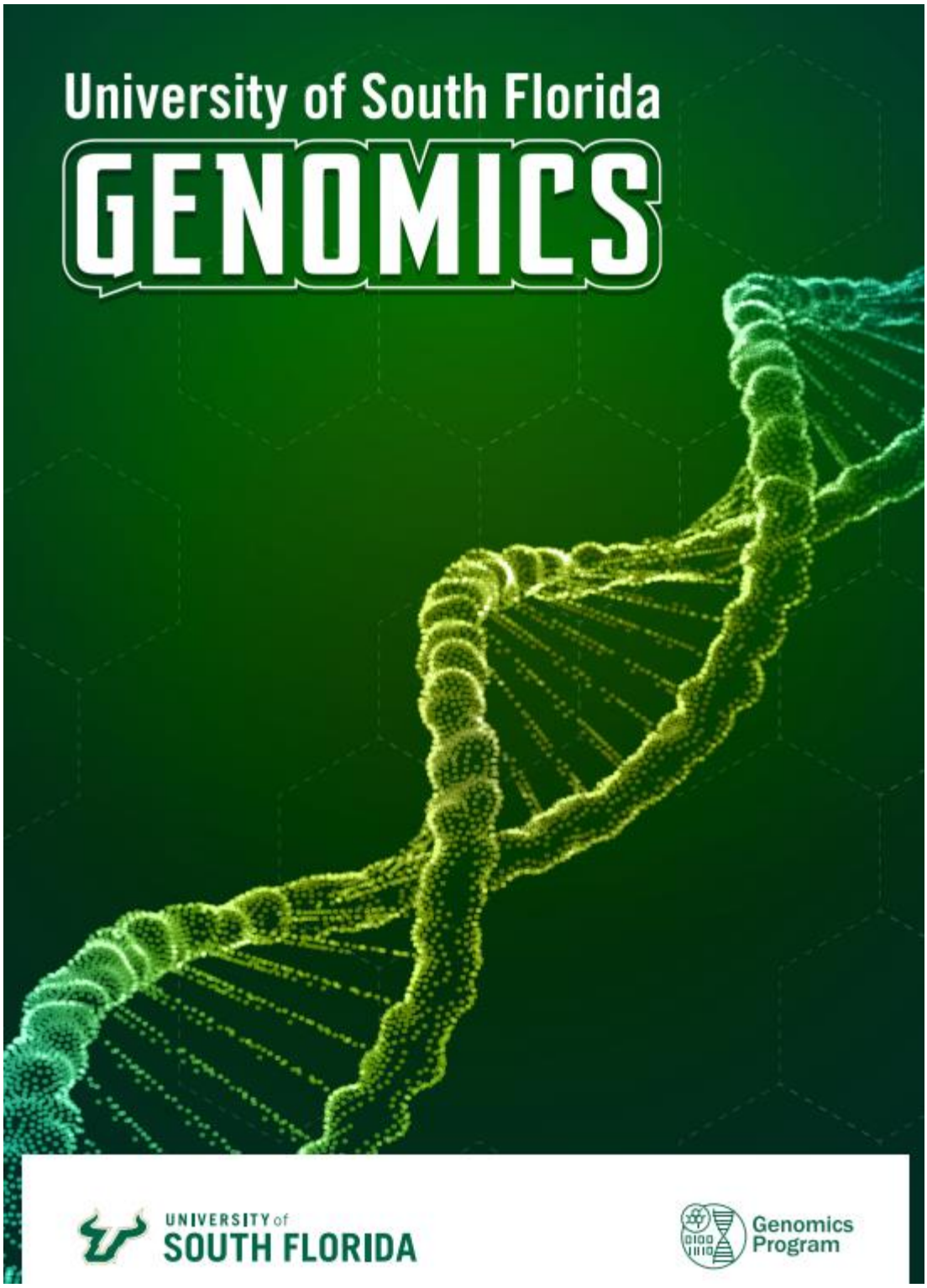
In association with the International Conference on Intelligent Biology and Medicine (ICIBM 2023) and The Genomics and Translational Bioinformatics working group of the American Medical Informatics Association, *Patterns* will publish a special collection on advances in translational bioinformatics guest edited by Panayiotis V. Benos, Zhongming Zhao, and Kai Wang. Papers accepted to ICIBM 2023 that become part of this collection will receive a 25% Article Publishing Charge (APC) discount.

Learn more about *Patterns*: https://www.cell.com/patterns/home

https://onefloridaconsortium.org/

OneFlorida+ Clinical Research Network

The OneFlorida+ Clinical Research Network is a collaboration among researchers, clinicians and patients in Florida, Georgia and Alabama to create an enduring infrastructure for a wide range of health research, including pragmatic clinical trials, comparative effectiveness research, implementation science studies, observational research, and cohort discovery. Network partners include the University of Florida, Florida State University, the University of Miami, the University of South Florida, Emory University in Atlanta, and the University of Alabama at Birmingham, along with the six universities' affiliated health systems and practices. Other partners include AdventHealth (Orlando), Tallahassee Memorial HealthCare, Tampa General Hospital, Bond Community Health (Tallahassee), Community Health IT (Kennedy Space Center), Nicklaus Children's Hospital (Miami), Capital Health Plan (Tallahassee), Bendcare (Boca Raton, Florida) and the Florida Agency for Health Care Administration, which oversees the Florida Medicaid Program. OneFlorida+ partners are committed to conducting stakeholder-engaged research in partnership with health systems, clinicians, patients, payers, policymakers and communities. Our partners strive for efficiency, offering multiple approaches and tools to facilitate healthcare research for today's complex health issues. OneFlorida+ network partners also ensure that lessons from research conducted in the region's diverse settings are systematically captured and translated back into improved health, health care and health policy for residents throughout the southeastern United States.


University of South Florida
GENOMICS
UNIVERSITY of
SOUTH FLORIDA
Genomics
Program