ICIBM 2022

International Conference on Intelligent Biology and Medicine

Hosted by

The International Association for Intelligent Biology and Medicine (IAIBM)

August 7-9, 2022
Philadelphia, PA, USA

2022 International Conference on
Intelligent Biology and Medicine
(ICIBM 2022)

August 7-9, 2022
Philadelphia, PA, USA

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

Table of content

Welcome to ICIBM 2022!

On behalf of all our conference committees and organizers, we welcome you to the 2022 International Conference on Intelligent Biology and Medicine (ICIBM 2022). 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 at a rapid pace and continue to have a strong impact in scientific research and medical innovations. With this in mind, we are proud to have built on 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. Ludmil Alexandrov, Xihong Lin, Smita Krishnaswamy, and Mona Singh. Throughout the conference, we will also feature six eminent scholar speakers, Drs. Feixiong Cheng, Jiang Qian, Yuan Luo, Li Shen, Derrick Scott, and Lana Garmire. These researchers are world-renowned experts and we are privileged to host their talks at ICIBM 2022. We will also be hosting four workshops on the first day of the conference. In addition, talks will be given from faculty members, postdoctoral fellows, PhD 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 2022's program. We'd like to extend our thanks to our sponsors for making this event possible, including the National Science Foundation, 10x Genomics, ClinChoice, BGI Americas, and Vizgen.

Last but not least, our sincerest thanks to members of all our ICIBM 2022 committees, and to our volunteers for their valuable efforts. Their dedication to making ICIBM 2022 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 conference that is thought provoking, fosters collaboration and innovation, and is enjoyable for all of our attendees. Thank you for attending ICIBM 2022. We look forward to your participation in all our conference has to offer!

Sincerely,

Jinchuan Xing, PhD
Program co-Chair
Associate Professor,
Department of Genetics & Human Genetics
Institute of New Jersey
Rutgers, The State University of New Jersey

Yan Guo, PhD
Program co-Chair
Associate Professor,
Department of Internal Medicine, Division of Molecular Medicine
University of New Mexico

Kai Wang, PhD
General co-Chair
Associate Professor,
Raymond G. Perelman Center for Cellular and Molecular Therapeutics & Department of Pathology,
Children's Hospital of Philadelphia

Zhongming Zhao, PhD
General co-Chair
Professor and Director,
Center for Precision Health
School of Biomedical Informatics
UTHealth, Houston

ACKNOWLEDGEMENTS

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

Steering Committee
Huanmei Wu, Temple University
Yidong Chen, University of Texas Health Science Center at San Antonio
Xiaohua Tony Hu, Drexel University
Kun Huang, Indiana University
Jason Moore, University of Pennsylvania, USA
Yi Xing, University of Pennsylvania, USA
Jennifer Ibrahim, Temple University
Jake Chen, University of Alabama at Birmingham

Program Chairs
Jinchuan Xing, Rutgers University, USA
Yan Guo, University of New Mexico, USA

Program Committee
Xiao Chang, Children's Hospital of Philadelphia, USA
Li Chen, Indiana University School of Medicine, USA
Junje Chen, Harbin Institute of Technology, Shenzhen, China
Bin Chen, Michigan State University, USA
Jianlin Cheng, University of Missouri Columbia, USA
Yong Cheng, St. Jude Children's Research Hospital, USA
Zechen Cheng, University of Alabama at Birmingham, USA
Lei Du, Northwestern Polytechnical University, USA
Shiaofen Fang, Indiana University-Purdue University Indianapolis, USA
Jun-Tao Guo, UNC Charlotte, USA
Leng Han, Texas A&M University Health Science Center, USA
Matthew Hayes, Xavier University of Louisiana, USA
Ruifeng Hu, Harvard University, USA
Ting Hu, Queens University, Canada
Tao Huang, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China
Tao Frank Huang, University of Cincinnati, USA
Weichuan Huang, EPA, USA
Zhi-Liang Ji, Xiamen University, China
Peilin Jia, Univsity of Texas Health Science Houston, USA
Limin Jiang, Chinese Academy of Sciences, China
Yufang Jin, The University of Texas at San Antonio, USA
Aman Cha Kaushik, School of Biotechnology, Gautam Buddha University, India
Fuhai Li, The Methodist Hospital, Weill Medical College of Cornell University, USA

Shuai Che Li, City University of Hong Kong, China
Aimin Li, Xi'an University of Technology, China
Tao Li, Nankai University, China
Li Liao, University of Delaware, USA
Kefei Liu, University of Pennsylvania, USA
Qian Liu, University of Nevada, Las Vegas, USA
Xiaoming Liu, University of South Florida, USA
Yaping Liu, Cincinnati Children's Hospital Medical Center, USA
Shuang Luan, University of New Mexico, USA
Tianle Ma, Oakland University, USA
Mirjana Maletic-Savatic, Baylor College of Medicine, USA
Huaiyu Mi, University of Southern California, USA
Danielle Mowery, University of Pennsylvania, USA
Kwangsik Nho, Indiana University, USA
Hong Qin, University of Tennessee Chattanooga, USA
Maciej Pietrzak, Ohio State University, USA
Li Shen, University of Pennsylvania, USA
Yang Shen, Texas A&M University, USA
Yufeng Shen, Columbia University, USA
Xinghua Shi, Temple University, USA
Zhifu Sun, Mayo Clinic, USA
Wing-Kin Sung, National University of Singapore, Singapore
Haixu Tang, Indiana University School of Medicine, USA
Manabu Torii, Kaiser Permanente, USA
Fuchiang Rich Tsui, Children's Hospital of Philadelphia/University of Pennsylvania, USA
Jun Wan, Indiana University School of Medicine, USA
Junab Wang, Radium Hospital, USA
Daifeng Wang, University of Wisconsin-Madison, USA
Jiayin Wang, Xi'an Jiaotong University, China
Chaochun Wei, Shanghai Jiao Tong University, China
Yingying Wei, The Chinese University of Hong Kong, China
Huanmei Wu, Temple University, USA
Junfeng Xia, Anhui University, China
Lei, Xie City University of New York, USA
Min, Xu Carnegie Mellon University, USA
Jianhua Xuan, Virginia Tech, USA
Yu Xue, Huazhong University of Science and Technology, China
Jingwen Yan, IUPUI, USA
Rendong Yang, University of Minnesota Twin Cities, USA
Xiaohui Yao, University of Pennsylvania, USA
Hui Yu, Vanderbilt University, USA
Lanjing Zhang, Rutgers University, USA
Wei Zhang, University of Central Florida, USA
Han Zhang, Nankai University, China
Shaojie Zhang, University of Central Florida, USA
Jim Zheng, Univsity of Texas Health Science Houston, USA

Cuncong Zhong, University of Kansas, USA
Yunyun Zhou, Children's Hospital of Philadelphia, USA

Publication Committee
James Cai, Co-Chair, Texas A&M University, USA
Hong Qin, Co-Chair, University of Tennessee at Chattanooga, USA

Workshop/Tutorial Committee
Yulin Dai, Co-Chair, University of Texas Health Science Center at Houston, USA
Daifeng Wang, Co-Chair, University of Wisconsin-Madison, USA

Publicity Committee
Leng Han, Co-Chair, Texas A&M University Health Science Center, USA
Kaifu Chen, Co-Chair, Harvard Medical School, USA

Award Committee
Christina Bergey, Co-Chair, Rutgers University, USA
Frank Huang, Co-Chair, University of Cincinnati College of Medicine, USA

Trainee Committee
Daniel Osorio Hurtado, Co-Chair, University of Texas at Austin, USA
Vanessa Nobles, Co-Chair, University of Arizona, USA

Local Organization Committee
Wanding Zhou, Co-Chair, Children's Hospital of Philadelphia, USA

Website Chair
Krishnamurthy Subramanian, Rutgers University, USA

International Conference on Intelligent Biology and Medicine Program

Sunday, Aug 7th

<table>
<tr><td>11:00 AM-7:30 PM</td><td colspan="3">Registration Open</td></tr>
<tr><td colspan="4">CONCURRENT WORKSHOPS/TUTORIALS</td></tr>
<tr><td></td><td>Room: Grand Ballroom</td><td>Room: Logan</td><td>Room: Rittenhouse</td></tr>
<tr><td>1:00-2:45 PM</td><td>Dr. Martin Renqiang Min, NEC-Labs

Deep Learning for Precision Immunotherapy</td><td rowspan="3">10:30 AM - 4:45 PM
Dr. Huanmei Wu, Temple University (Chair)
Dr. Hongfang Liu, Mayo Clinic
Dr. Qi Wang, University of South Carolina
Dr. Jun Deng, Yale University
Dr. Hua Xu, University of Texas (Houston)
Dr. Rui Zhang, University of Minnesota - Twin Cities
Dr. Jay Patel, Temple University (Publication Chair)
Dr. Yonghui Wu, University of Florida
Dr. Yanshan Wang, University of Pittsburgh
Dr. Derrick Scott, Delaware State University (Eminent Scholar)
Jennifer Ibrahim, Temple University College of Public Health
Dr. Omar Martinez, University of Central Florida
Dr. Lixia Yao, Merck
Dr. David Fleece, Temple University School of Medicine

International Workshop on Translating Scientific Discoveries into Action (I2A)</td><td rowspan="3">1:00 PM - 4:00 PM
Dr. Yichuan Zhao, Georgia State University
Dr. Tian Tian, CHOP
Dr. Jung-Ying Tzeng, North Carolina State University
Dr. Xinlei (Sherry) Wang, Southern Methodist University

Big data with statistical approaches</td></tr>
<tr><td>2:45-3:00 PM</td><td>Coffee Break</td></tr>
<tr><td>3:00-4:45 PM</td><td>Dr. Daifeng Wang, University of Wisconsin-Madison
Dr. Anru Zhang, Duke University

Multimodal data integration and analysis</td></tr>
<tr><td>4:45-5:00 PM</td><td colspan="3">Break</td></tr>
<tr><td>5:00-5:40 PM</td><td colspan="3">Keynote Lecture (Room: Grand Ballroom)

Ludmil Alexandrov, Ph.D.

Anthology of unusual patterns of somatic mutations in cancer genomes</td></tr>
<tr><td>5:40-5:45 PM</td><td colspan="3">Break</td></tr>
<tr><td>5:45-7:30 PM</td><td colspan="3">Poster Session
Abstract ID: 34, 50, 52, 53, 54, 58, 60, 61, 66, 68, 69, 71, 80, 83, 86, 90, 91, 96, 97, 103, 107, 108, 114, 116</td></tr>
</table>

Monday, Aug 8th

7:30 AM - 5:30 PM	**Registration Open**		
8:30 AM - 8:40 AM	**Opening Remarks (Jinchuan Xing, Yan Guo)**		
8:40 AM - 9:20 AM	**Keynote Lecture (Room: Main)** **Mona Singh, Ph.D.** ***Deciphering cellular interaction networks: From normal functioning to disease***		
9:20 AM - 9:30 AM	***Break for parallel sessions***		
CONCURRENT SESSIONS			
	Room: Grand Ballroom	**Room: Logan**	**Room: Rittenhouse**
	Integrative multi-omics view of disease: from genetic variations to epigenetic dysregulation *Session Chair:* Kaifu Chen	**Machine learning in Cancer** *Session Chair:* Shaolei Teng Marek Kimmel	**Informatics in team science: to lead, support, and educate** *Session Chair:* Li Liu Gangqing Hu
9:30 AM - 9:50 AM	**Microbiome and tumor microenvironment crosstalk in the gastric cancer** Chao Zhang	**Eminent Scholar** **Luo Yuan** ***Machine Learning on Multi-Modal Healthcare Data***	**The WVU Bioinformatics Core: Empowering Biomedical Research through Team Science in 'Almost Heaven' West Virginia** Gangqing (Michael) Hu
9:50 AM - 10:10 AM	**Specific role of CTCF in oncogenic transcriptional dysregulation** Chongzhi Zang	**Risk Stratification for Breast Cancer Patient by Simultaneous Learning of Molecular Subtype and Survival Outcome Using Genetic Algorithm-Based Gene Set Selection** Bonil Koo, Dohoon Lee, Sangseon Lee, Inyoung Sung and Sun Kim	**Team science in a large medical institution: the tradition of "the needs of the patient come first"** Zhifu Sun
10:10 AM -10:30 AM	**Linking genetic variants to kidney disease via the epigenome** Hongbo Liu	**A Novel Bayesian Framework Infers Driver Activation States and Reveals Pathway-oriented**	**NGS-related bioinformatics in Academy and Industry** Yaping Feng

		Molecular Subtypes in Head and Neck Cancer Zhengping Liu, Chunhui Cai, Xiaojun Ma, Jinling Liu, Lujia Chen, Vivian Lui, Gregory Cooper and Xinghua Lu	
10:30 AM -10:50 AM	**RNA m6A landscape reveals strong regulatability of tumor suppressor expression** Kaifu Chen	**Robust personalized classifier improves the prediction of breast cancer metastasis (remote)** Nahim Adnan and Jianhua Ruan	**Education in bioinformatics: how to position and balance?** Jingwen Yan
10:50 AM -11:05 AM	***Coffee Break***		
11:05 AM - 11:25 AM	**Harnessing big data to characterize toxicity of immunotherapy** Leng Han	**High Resolution Cell Type Deconvolution Reveals Cell Type Specific Molecular Mechanism of Cancer Radioresistance** Xiao Sun, Min Zhu, Xiaoyi Fei and Xueling Li	**Biomedical informatics training in diverse environments** Li Liu
11:25 AM -11:45 AM	**Cobind: quantify the magnitude of genomic collocation** Liguo Wang	**Modeling the relationship between gene expression and mutational signature** Limin Jiang, Hui Yu and Yan Guo	**Panel Discussion**
11:45 AM -12:05 PM	**Dissect human diseases with high throughput single cell nanopore sequencing** Ruli Gao	**Flash Talk (6 x 5mins)** 11:45 AM -12:15 PM	**Flash Talk (6 x 5mins)** 11:45 AM -12:15 PM
		Structure-enhanced Deep Meta-learning Predicts Uncharted Chemical-Protein Interactions on a Genome-scale Tian Cai, Li Xie, Shuo Zhang, Muge Chen and Lei Xie	**Identification of genetic loci associated with the risk of aneuploidy with maternal-origin using PGT-A sequences** Siqi Sun, Aishee Bag, Daniel Ariad, Mary Haywood, Mandy Katz-Jaffe, Rajiv McCoy, Karen Schindler and Jinchuan Xing
		Deep Learning Prediction of Chemical-induced Dose-	**An R Shiny app for systematically integrating**

		Dependent and Context-Specific Multiplex Phenotype Responses and Its Application to Personalized Alzheimer's Disease Drug Repurposing You Wu, Qiao Liu, Yue Qiu and Lei Xie	**genetic and pharmacologic cancer dependency maps** Tapsya Nayak, Li-Ju Wang, Michael Ning, Yufei Huang, Yu-Chiao Chiu and Yidong Chen
		SPCount: Advances in deep taxonomy alignments for extracellular small RNAs Quanhu Sheng, Marisol Ramirez, Ryan Allen, Qi Liu, Michelle Ormseth, Kasey Vickers and Yu Shyr	**dRFEtools: Dynamic recursive feature elimination for omics** Kynon Jade Benjamin, Tarun Katipalli and Apuã Paquola
12:05 PM - 12:25 PM	**Extend the health-span: decipher the epigenome of aging and leukemogenesis** Sheng Li	**Genome-wide analysis on model derived binge-eating disorder phenotype identifies the first three risk loci and implicates iron metabolism** David Burstein, Trevor Griffen, Karen Therrien, Jaroslav Bendl, Sanan Venkatesh, Biao Zeng, Amirhossein Modabbernia, Pengfei Dong, Deepika Mathur, Gabriel Hoffman, Robyn Sysko, Tom Hildebrandt, Georgios Voloudakis and Panos Roussos	**Genomic Data Augmentation Based on Few-shot Generative Domain Adaptation** Chen Song, Emily Thyrum and Xinghua Shi
		A new player into the cellular heterogeneity: the flexible and mobile extrachromosomal circular DNA Jiajinlong Kang, Yulin Dai, Jinze Li, Huihui Fan and Zhongming Zhao	**Prediction of Pathological Stages in Prostate Cancer Using Graph Attention Networks** Wenkang Zhan, Chen Song, Zhengkang Fan and Xinghua Shi
		Fifty-one novel, replicated loci identified in genome-wide association study of polyunsaturated and	**Prediction of return of spontaneous circulation using physiological waveforms during**

		monounsaturated fatty acids in 124,024 European individuals Michael Francis, Yitang Sun, Huifang Xu, James Brenna and Kaixiong Ye	**cardiopulmonary resuscitation in a pediatric experimental model of cardiac arrest: a machine learning pilot study** Luiz Eduardo Silva, Lingyun Shi, Tiffany Ko, Hunter Gaudio, Vivek Padmanabhan, Ryan Morgan, Julia Slovis, Rodrigo Forti, Sarah Morton, Yuxi Lin, Gerard Laurent, Jake Breimann, Bo Yun, Nicolina Ranieri, Madison Bowe, Wesley Baker, Todd Kilbaugh and Fuchiang Tsui
12:25 PM - 1:40 PM	***Lunch Break***		
1:40 PM - 2:20 PM	**Keynote Lecture (Room: Grand Ballroom)** **Smita Krishnaswamy, Ph.D.** ***Graph-based signal processing and machine learning for extracting structure from biomedical data***		
2:20 PM - 2:30 PM	***Break for parallel sessions***		
CONCURRENT SESSIONS			
	Room: Grand Ballroom	**Room: Logan**	**Room: Rittenhouse**
	Cancer informatics Session Chair: Xiaoming Liu Jin Lu	**Network approaches in biomedical research** Session Chairs: Lijun Cheng Xiaoyi Raymond Gao	**Artificial Intelligence on Big Data: Promise for Early-stage Trainees** Session Chairs: Yufang Jin Yu-Chiao Chiu Chi Zhang Yongsheng Bai
2:30 PM - 2:50 PM	**Eminent Scholar** **Jiang Qian** ***Feedback loops in cell-cell communication***	**Eminent Scholar** **Feixiong Cheng** ***Harnessing Endophenotypes and Network Medicine for Discovery of Pathobiology and Drug Repurposing in Alzheimer's Disease***	**5 Mins Talks** **White blood cells and obesity: A Mendelian randomization study** James Yang, Yitang Sun and Kaixiong Ye

			CoMutDB: The Landscape of Somatic Mutation Co-occurrence in Cancers Chaoyi Troy Zhang and Yan Guo **Identifying Putative Causal Links between MicroRNAs and Severe COVID-19 Using Mendelian Randomization** Chang Li, Aurora Wu, Kevin Song, Jeslyn Gao, Eric Huang, Yongsheng Bai and Xiaoming Liu
2:50 PM - 3:10 PM	**CellCallEXT: analysis of ligand–receptor and transcription factor activities in cell–cell communications of tumor immune microenvironment** Shouguo Gao, Xingmin Feng, Zhijie Wu, Sachiko Kajigaya and Neal Young	**NetCellMatch: Multiscale Network-Based Matching of Cancer Cell Lines to Patients Using Graphical Wavelets** Neel Desai, Jeffrey Morris and Veera Baladandayuthapani	**A Deep Learning Model for Ancestry Estimation with Craniometric Measurements** Kevin Ma and Xiaoming Liu **Condition-specific Gene Co-expression Network Analysis Reveals Copy Number Variations Associated with KRAS Mutation Status in Colon Cancer** Shaoyang Huang and Chi Zhang **Computational modeling of cell type specific metabolic rate of glucose flow and glutaminolysis in cancer microenvironment** Grace Yang, Kevin Hu, Alex Lu, Shaoyang Huang, Pengtao Dang, Haiqi Zhu, Sha Cao and Chi Zhang
3:10 PM - 3:30 PM	**Identification of immuno-targeted combination therapies using explanatory subgroup discovery for cancer patients with**	**PPIGCF: A Protein-Protein Interaction Based Gene Correlation Filter for Optimal Gene Selection**	**Computational modeling of cell type specific metabolic rate of Branched Chain Amino Acids** Kevin Hu, Alex Lu, Shaoyang Huang, Grace Yang, Haiqi

	EGFR wild-type (WT) gene Olha Kholod, William Basket, Danlu Liu, Jonathan Mitchem, Jussuf Kaifi, Laura Dooley and Chi-Ren Shyu	Soumen Kumar Pati, Manan Kumar Gupta, Ayan Banerjee, Saurav Mallik and Zhongming Zhao	Zhu, Pengtao Dang, Sha Cao and Chi Zhang **Interactive Data Collection Using CARLA and OpenCDA for Reinforcement Learning** Alan Chen, Joseph Clemmons, Umar Jamil, Ashley Land, Sara Ahmed Yu-Fang Jin **Optimal Charging Strategy for Electric Vehicles** R.J. Alva, Albert Zhang, Eugenia Cadete, Yu-Fang Jin and Sara Ahmed
3:30 PM - 3:50 PM	**Spatial transcriptomic analysis reveals associations between genes and cellular topology in breast and prostate cancers** Lujain Alsaleh, Chen Li, Justin Couetil, Kun Huang, Jie Zhang, Chao Chen and Travis Johnson	**Cancer Classification from Gene Expression Using Graph Attention Network (remote)** Sheikh Muhammad Saiful Islam, Ziqian Xie and Degui Zhi	**Edge-Based AI Navigation System for Automated Wheelchairs** Jeffrey Wang, Kevin Liu and Yu-Fang Jin **Detecting Unattended Baby in Car-seat as a New Vehicle Safety Feature** Jerry Guo, William Xiao and Yu-Fang Jin
3:50 PM - 4:00 PM	***Coffee Break***		
4:00 PM - 4:20 PM	**Clone phylogenetics reveals metastatic tumor migrations, maps, and models** Antonia Chroni, Sayaka Miura, Lauren Hamilton, Tracy Vu, Stephen Gaffney, Vivian Aly, Sajjad Karim, Maxwell Sanderford, Jeffrey Townsend and Sudhir Kumar	**Identify Potential Driver Genes for PAX-FOXO1 Fusion-Negative Rhabdomyosarcoma Through Frequent Gene Co-expression Network Mining** Xiaohui Zhan, Yusong Liu, Asha Jacob Jannu, Shaoyang Huang, Bo Ye, Wei Wei, Pankita H. Pandya, Xiufen Ye, Karen E. Pollok, Jamie L. Renbarger, Kun Huang and Jie Zhang	**10 Mins Talks** **Decentralized Collision-Free Trajectory Planning for Autonomous Vehicles Using Reinforcement Learning** Joseph Clemmons, Umar Jamil, Alan Chen, Ashley Land, Sara Ahmed and Yu-Fang Jin **Developing a Sustainable Low-maintenance Traffic Crash Risk Notification System**

			Seth Klupka, Tulan Sampath Bandara, Paul Morton, Mimi Xie and Yu-Fang Jin
4:20 PM - 4:40 PM	**Genomic variants disrupt miRNA-mRNA regulation** Ellie Xi, Judy Bai, Klaira Zhang, Hui Yu and Yan Guo	**A Boolean network analysis of Parkinson's Disease genes and mitochondrial metabolic factors influencing Oxidative phosphorylation (remote)** Baby Kumari, Md. Zainul Ali and Pankaj Singh Dholaniya	**Transfer Learning for Cancer Survival Prediction using Gene Expression Data** Ariel Lee, Jacob Buckle, Ricardo Ramirez and Yufang Jin **Analysis for Traffic Crash Severity in Texas Using CRIS Database** Shuyu He, Taylor Jones, Tulan Sampath Bandara, Seth Klupka, Sara Ahmed and Mimi Xie
4:40 PM - 5:10 PM	**Flash Talk (6x5m)**	**Flash Talk (6x5m)**	**Cell-Type Identification with Single Cell RNA-Sequencing Data for Temporomandibular Joint Disorders** Mostafa Malmir, Savannah Lopez, Marlene Luo, Karen Lindquist, Sergei Belugin, Arman Akopian, Yidong Chen, Jinyan Li and Yu-Fang Jin **Retrieving Knowledge of Molecular Mechanisms from Literature Titles via an Event Extraction Approach** David Spellman, Jason Xiaotian Dou, Aaron Fangzheng Wu and Yufei Huang **Mapping Imaging Genetics Associations at Multiple Scales: A Study of Cortical Thickness Phenotypes for Alzheimer's Disease**
	PRIME Evolutionary Imputation (PREI) Hannah Kim and Sergei Pond	**Identifying drug hepatotoxicity mechanisms using high-throughput concentration-dependent toxicity data and toxicokinetic modeling** Daniel Russo, Lauren Aleksunes and Hao Zhu	
	Systems Pharmacogenomic Framework Identifies Associations between Gut-microbiota Metabolites and GPCRome in Alzheimer's disease Yunguang Qiu, Yuan Hou, Yadi Zhou, Jielin Xu, James B. Leverenz, Andrew A. Pieper, Jeffrey Cummings and Feixiong Cheng	**Benchmark study of similarity measures from query phenotypic abnormalities to diseases based on the human phenotype ontology** Yu Hu and Kai Wang	
	Tensor-Based Multi-Modality Multi-Target Regression for	**Functional Impact of Copy Number Variants in Autism Spectrum Disorder and Related Disorders**	

	Alzheimer's Disease Diagnosis Jun Yu, Yong Chen, Li Shen and Lifang He	Rohan Alibutud, Vaidhyanathan Mahaganapthy, Xiaolong Cao, Marco Azaro, Christine Gwin, Sherri Wilson, Steven Buyske, Christopher Bartlett, Judy Flax, Linda Brzustowicz and Jinchuan Xing	Kevin Shen, Manu Shivakumar, Dokyoon Kim **Round Table Discussion**
	Identifying Chronic Tic Disorder subtypes using clinical diagnostic data Subramanian Krishnamurthy, Tourette International Collaborative Genetics and Jinchuan Xing	**Immuno-therapy-induced gene signatures in clear cell renal cell carcinoma (remote)** Ye-Lin Son, Huihui Fan and Zhongming Zhao	
	Sequences of Events from the Electronic Medical Record and the Onset of Infection Caitlin Coombes, Kevin Coombes and Naleef Fareed	**Synchronized decoding of functional capacities and compositions of metagenomes in a sweep** Daniel Roush, Daniel Hakim, Antonio González, George Armstrong, Justin Shaffer, Daniel McDonald, Rob Knight and Qiyun Zhu	
	A pan-cancer analysis and identification of glucose-6-phosphate dehydrogenase (G6PD) inhibitors by cheminformatics approach Madhu Sudhana Saddala and Jiang Qian	**A novel water aware QM-MM hybrid method for macromolecule drug discovery and bio-active conformational prediction (remote)** Yuxin Xie, Wanting Chen and Shengpei Chen	
5:10 PM - 5:30 PM	***Award presentation***		
6:30 PM	**BANQUET**		

Tuesday, Aug 9th

7:30 AM	**Registration Open**		
8:30 AM - 9:10 AM	**Keynote Lecture (Grand Ballroom)** **Xihong Lin, Ph.D.** ***Scalable Integrative Analysis of Large-Scale Biobanks, Whole Genome Sequencing Studies and Functional Multi-Omic Data***		
9:10 AM - 9:20 AM	***Break for parallel sessions***		
CONCURRENT SESSIONS			
	Room: Grand Ballroom	**Room: Logan**	**Room: Rittenhouse**
	Application of machine learning techniques in genetics and genomics Session Chair: Shizhong Han Bingshan Li	**Bioinformatics, Genomics** Session Chair: Kaixiong Ye Qiyun Zhu	**Single-cell Omics** Session Chair: Quanhu Sheng Travis S. Johnson
9:30 AM - 9:50 AM	**Pathogenicity prediction for nonsynonymous SNVs and non-frameshift Indels** Xiaoming Liu	**Eminent Scholar** **Li Shen** ***Integrating imaging and genomics data for gene discovery in Alzheimer's disease***	**Eminent Scholar** **Lana Garmire** ***Pushing single cell data science towards clinical applications***
9:50 AM - 10:10 AM	**Transformer-based unsupervised learning for spatial transcriptomic analysis** Wei Chen (Chongyue Zhao)	**NSP1 inhibition effects on Human 40s ribosomes compared to intermediate carrier animals** Bijan Bambai, Mitra Salehi, Zarrin Minuchehr and Afagh Bapirzadeh	**Time-varying gene expression network analysis reveals conserved transition states in hematopoietic differentiation between human and mouse** Shouguo Gao, Ye Chen, Zhijie Wu, Sachiko Kajigaya, Xujing Wang and Neal Young
10:10 AM - 10:30 AM	**Assessing Tissue-specific Functional Effects of Non-coding Variants with Deep Learning**	**Prediction of the effects of missense mutations on human Myeloperoxidase protein stability using in**	**Robust Augmenting Single-cell RNA-seq with Surface Protein Levels using Geneset Deep Learning and Transfer Learning**

	Bingshan Li	**silico saturation mutagenesis** Adebiyi Sobitan, William Edwards, Md Shah Jalal, Kolawole Ayanfe, Hemayet Ullah, Atanu Duttaroy, Jiang Li and Shaolei Teng	Md Musaddaqul Hasib, Tinghe Zhang, Jianqiu Zhang, Shou-Jiang Gao and Yufei Huang
10:30 AM - 10:50 AM	**CNN Algorithms for Disease Classification and Biomarker Discovery** Xiangning Chen	**DelInsCaller: An Efficient Algorithm for Identifying Delins from Long-Read Sequencing Data with High-Level of Sequencing Errors (remote)** Shenjie Wang, Xuanping Zhang, Geng Qiang and Jiayin Wang	**GenKI: a variational graph autoencoder based virtual knockout tool for gene function predictions via single-cell gene regulatory network** Yongjian Yang, Guanxun Li, Yan Zhong, Qian Xu and James Cai
10:50 AM - 11:05 AM	***Coffee/Tea Break***		
11:05 AM - 11:25 AM	**Genome-wide cell-free DNA fragmentation as a biomarker for early detection of cancer** Stephen Cristiano	**microRNA and microRNA target variants associated with autism spectrum disorder and related disorders** Anthony Wong, Anbo Zhou, Xiaolong Cao, Vaidhyanathan Mahaganapathy, Marco Azaro, Christine Gwin, Sherri Wilson, Steve Buyske, Christopher W. Bartlett, Judy F. Flax, Linda Brzustowicz and Jinchuan Xing	**EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps** Xiaotao Wang, Yu Luan and Feng Yue
11:25 AM - 11:45 AM	**Deep learning predicts DNA methylation regulatory variants in the human brain and elucidates the genetics of psychiatric disorders** Shizhong Han	**TransModDNA: Transformer-based DNA Methylation detection on ionic signals from Oxford Nanopore sequencing data**	**Estimation of clonal dynamics of lung cancers based on DNA sequencing data: Potential impact on detection and cure** Marek Kimmel, Khanh Dinh and Andrew Koval

		Xiuquan Wang, Mian Umair Ahsan, Yunyun Zhou and Kai Wang	
11:45 AM - 12:05 PM	**Machine learning approaches to enhance gene expression prediction integrating eQTLs with 3D genomes and epigenetic data** Chachrit Khunsriraksakul	**mintRULS: Prediction of miRNA-mRNA target site interactions using regularized least square method** Sushil Shakyawar, Siddesh Southekal and Babu Guda	**Combination of serum and plasma biomarkers could improve prediction performance for Alzheimer's Disease (remote)** Fan Zhang, Melissa Petersen, Leigh Johnson, James Hall and Sid O'Bryant
12:05 PM - 1:30 PM	***Lunch Break***		
1:30 PM - 1:40 PM	***Coffee Break***		
CONCURRENT SESSIONS			
	Room: Grand Ballroom	**Room: Logan**	**Room: Rittenhouse**
	Machine learning in biomedical research Session Chair: Fuhai Li	**COVID-19 Informatics** Session Chair: Yufeng Shen	**Medical informatics** Session Chair: Luiz Eduardo Silva
1:40 PM - 2:00 PM	**Secure and Efficient Implementation of Facial Emotion Detection for Smart Patient Monitoring System (remote)** Kh Shahriya Zaman and Md Mamun Bin Ibne Reaz	**Drug-Target Network Study Reveals the Core Target-protein Interactions of Various COVID-19 Treatments** Yulin Dai, Hui Yu, Qiheng Yan, Bingrui Li, Andi Liu, Wendao Liu, Xiaoqian Jiang, Yejin Km, Yan Guo and Zhongming Zhao	**Classifying Refugee Status Using Common Features in EMR** Malia Morrison, Vanessa Nobles, Crista Johnson-Agbakwu, Celeste Bailey and Li Liu
2:00 PM - 2:20 PM	**DeepG2P: predicting protein abundance from mRNA expression** Hui-Mei Tsai, Tzu-Hung Hsiao, Yu-Chiao Chiu, Yufei Huang, Yidong Chen and Eric Y. Chuang	**Integrated analysis of bulk RNA-seq and single-cell RNA-seq unravels the influences of SARS-CoV-2 infections to cancer patients**	**Using machine learning to assess the range of services provided by family physicians in Ontario, Canada**

		Yu Chen, Yujia Qin, Yuanyuan Fu, Zitong Gao and Youping Deng	Arunim Garg, David Savage, Salimur Choudhury and Vijay Mago
2:20 PM - 2:40 PM	**Transformer for Gene Expression Modeling (T-GEM): An interpretable deep learning model for gene expression-based phenotype predictions (remote)** Ting-He Zhang, Md Musaddaqui Hasib, Yu-Chiao Chiu, Zhi-Feng Han, Yu-Fang Jin, Mario Flores, Yidong Chen and Yufei Huang	**Cell-Specific Gene Signature for COVID-19 Infection Severity Using single-cell RNA-seq analysis** Mario Flores, Karla Paniagua, Wenjian Huang, Ricardo Ramirez, Yidong Chen, Yufei Huang and Yu-Fang Jin	**Termviewer - A Web Application for Streamlined Human Phenotype Ontology (HPO) Tagging and Document Annotation** Anna Nixon, James M. Havrilla, Li Fang and Kai Wang
2:40 PM - 3:00 PM	**Interpretable Drug Synergy Prediction with Graph Neural Networks for Human-AI Collaboration in Healthcare** Zehao Dong, Heming Zhang, Yixin Chen and Fuhai Li		**Mining High-level Imaging Genetic Associations via Clustering AD Candidate Variants with Similar Brain Association Patterns** Ruiming Wu, Jingxuan Bao, Mansu Kim, Andrew Saykin, Jason Moore and Li Shen
3:00 PM - 3:10 PM	***Wrap-up and Closing Remarks***		

COVID-19 Guidelines:

We will follow the latest CDC and Philadelphia guidelines: CDC Guidelines on COVID-19; Philadelphia guidelines on COVID-19. Please keep practicing good hand hygiene and respiratory etiquette, and protect yourselves and avoid close contact based on personal risk factors laid out by the CDC.

Keynote Speaker
Ludmil Alexandrov, Ph.D.
Sunday, August 7, 2022
5:00 – 5:45 PM
Grand Ballroom

Bio

I am an enthusiastic early career scientist with an interdisciplinary training and a strong computational background. My interests lie in leveraging the information hidden in large-scale omics data for better understanding of the mutational processes causing human cancer, for identifying potential cancer prevention strategies, and for developing novel approaches for targeted cancer treatment.

My research has been focused on understanding mutational processes in human cancer through the use of mutational signatures. In 2013, I developed the first comprehensive map of the signatures of the mutational processes that cause somatic mutations in human cancer. This work was published in several well-regarded scientific journals and highlighted by the American Society of Clinical Oncology as a milestone in the fight against cancer. More recently, I mapped the signatures of the clock-like mutational processes operative in normal somatic cells, demonstrated that mutational signatures have the potential to be used for targeted cancer therapy, and identified the mutational signatures associated with tobacco smoking. My interests, expertise, and interpersonal skills are backed up by more than 100 scientific publications, prestigious national and international awards, as well as work experience in government laboratories, renowned universities, and leading consulting companies.

During the past few years, I have received multiple awards for my work on mutational signatures in human cancer. Most recently, I was awarded a 2020 NIEHS Outstanding New Environmental Scientist Award.

Title: Anthology of unusual patterns of somatic mutations in cancer genomes

Cancer is the most common human genetic disease. All cancers are caused by somatic mutations. These mutations may be the consequence of the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA, or defective DNA repair. In some cancer types, a substantial proportion of somatic

mutations are known to be generated by exogenous carcinogens, for example, tobacco smoking in lung cancers and ultraviolet light in skin cancers, or by abnormalities of DNA maintenance, for example, defective DNA mismatch repair in some colorectal cancers.

Each biological process causing mutations leaves a characteristic imprint on the genome of a cancer cell, termed, mutational signature. In this talk, I will present an anthology describing the clinical and scientific utility of mutational signatures across the spectrum of human cancers. First, in our stories of the past, I will describe mutational signatures and how they have been previously used for developing cancer prevention strategies and for targeted cancer treatment. Second, in the anecdotes of the present, I will show the recent analysis that mapped the repertoire of mutational signatures of clustered events and copy-number changes. Further, I will also present a novel machine learning approach for detecting homologous recombination deficiency. Lastly, in our dreams of the future, I will describe how deep learning AI approaches can be utilized for addressing inequalities in cancer diagnosis. This bouquet of mutational pattern stories reveals the diversity of mutational processes underlying the development of cancer and outlines the utility of mutational signatures for cancer treatment and cancer prevention.

Keynote Speaker
Mona Singh, Ph.D.
Monday, August 8, 2022
8:40 – 9:20 AM
Grand Ballroom

Bio

Mona Singh obtained her AB and SM degrees at Harvard University, and her PhD at MIT, all three in Computer Science. She did postdoctoral work at the Whitehead Institute for Biomedical Research. She has been on the faculty at Princeton since 1999, and currently she is the Wang Family Professor in Computer Science in the computer science department and the Lewis-Sigler Institute for Integrative Genomics. She received the Presidential Early Career Award for Scientists and Engineers (PECASE) in 2001, and is a Fellow of the International Society for Computational Biology and a Fellow for the Association for Computing Machinery. She is Editor-In-Chief of the Journal of Computational Biology. She has been program committee chair for several major computational biology conferences, including ISMB (2010), WABI (2010), ACM-BCB (2012), and RECOMB (2016), and has been Chair of the NIH Modeling and Analysis of Biological Systems Study Section (2012-2014).

Title: Deciphering cellular interaction networks: From normal functioning to disease

Each cell in our body accomplishes its functions via a complex network of molecular interactions. Knowledge of these networks are thus key to understanding cellular functioning (and, in the case of disease, malfunctioning). I will overview frameworks and algorithms that leverage interaction networks in order to gain a better understanding of diseases such as cancer.

Keynote Speaker
Smita Krishnaswamy, Ph.D.
Monday, August 8, 2022
1:40 – 2:20 PM
Grand Ballroom

Bio

Smita Krishnaswamy is an Associate Professor in the departments of Computer Science and Genetics at Yale University. She is part of the programs in Applied Mathematics, Computational Biology & Bioinformatics and Interdisciplinary Neuroscience. She is also affiliated with the Yale Center for Biomedical Data Science, Yale Cancer Center, Wu-Tsai Institute. Smita's lab works at the intersection of computer science, applied math, computational biology, and signal processing to develop representation-learning and deep learning methods that enable exploratory analysis, scientific inference and prediction from big biomedical datasets. She has applied her methods on datasets generated from single-cell sequencing, structural biology, biomedical imaging, brain activity recording, electronic health records on a wide variety of biological, cellular, and disease systems. Her techniques generally incorporate mathematical priors from graph spectral theory, manifold learning, signal processing, and topology into machine learning and deep learning frameworks, in order to denoise and model the underlying systems faithfully for predictive insight. Currently her methods are being widely used for data denoising, visualization, generative modeling, dynamics. modeling, comparative analysis and domain transfer.

Smita teaches several courses including: Deep Learning Theory and Applications, Unsupervised learning, and Geometric and Topological Methods in Machine Learning. Prior to joining Yale, Smita completed her postdoctoral training at Columbia University in the systems biology department where she focused on learning computational models of cellular signaling from single-cell mass cytometry data. She obtained her Ph.D. from EECS department at University of Michigan where her research focused on algorithms for automated synthesis and probabilistic verification of nanoscale logic circuits, winning an outstanding dissertation award from EDAA. Following her time in Michigan, Smita spent 2 years at IBM's TJ Watson Research Center as a researcher in the systems division where she worked on automated bug finding and error correction in logic. Smita's work over the years has won several awards

including the NSF CAREER Award, Sloan Faculty Fellowship, and Blavatnik fund for Innovation.

Title: Graph-based signal processing and machine learning for extracting structure from biomedical data

In this talk, I will show how to leverage data geometry and topology, embedded within modern machine learning frameworks, to understand complex high dimensional scientific data. First, I will show how graphs can model underlying manifolds from which data are sampled and how graph spectral tools such as diffusion operators and signal processing tools such as filters can shed light on characteristics of the underlying manifold including geodesic distances, density, and curvature. Next, I will show how to combine graph diffusion geometry with topology to extract multi-granular features from the data for predictive analysis. Then, I will move up from the local geometry of individual data points to the global geometry of complex objects like data clouds, using graph signal processing to derive representations of these entities and optimal transport for distances between them. Finally, I will demonstrate how two neural networks use geometric inductive biases for generation and inference: GRASSY (geometric scattering synthesis network) for generating new molecules and molecular fold trajectories, and TrajectoryNet for performing dynamic optimal transport between time-course samples to understand the dynamics of cell populations. Throughout the talk, I will include examples of how these methods shed light on the inner workings of biomedical and cellular systems including cancer, immunology and neuroscientific systems. I will finish by highlighting future directions of inquiry.

Keynote Speaker
Xihong Lin, Ph.D.
Tuesday, August 9, 2022
8:30 – 9:10 AM
Grand Ballroom

Bio

Xihong Lin, PhD is Professor and former Chair of the Department of Biostatistics, Coordinating Director of the Program in Quantitative Genomics at the Harvard T. H. Chan School of Public Health, and Professor of the Department of Statistics at the Faculty of Arts and Sciences of Harvard University, and Associate Member of the Broad Institute of MIT and Harvard. Dr. Lin's research interests lie in development and application of scalable statistical and machine learning methods for analysis of massive high-throughput data from genome, exposome and phenome, as well as complex epidemiological, biobank and health data. Dr. Lin received the MERIT Award (R37) (2007-2015) and the Outstanding Investigator Award (OIA) (R35) (2015-2022) from the National Cancer Institute (NCI). She is the contact PI of the Harvard Analysis Center of the NHGRI Genome Sequencing Program, and the multiple PI of one of the Predictive Modeling Centers of the NHGRI Impact of Genomic Variation on Function (IGVF) program. Dr. Lin is an elected member of the National Academy of Medicine. She has received several prestigious awards including the 2002 Mortimer Spiegelman Award from the American Public Health Association, the 2006 Presidents' Award of the Committee of Presidents of Statistical Societies (COPSS), and the 2022 Marvin Zelen Leadership in Statistical Science Award. She is an elected fellow of American Statistical Association, Institute of Mathematical Statistics, and International Statistical Institute. Dr. Lin is the former Chair of the COPSS (2010-2012) and a former member of the Committee of Applied and Theoretical Statistics of the National Academy of Science. She is the founding chair of the US Biostatistics Department Chair Group, and the founding co-chair of the Young Researcher Workshop of East-North American Region (ENAR) of International Biometric Society. She is the former Coordinating Editor of Biometrics and the founding co-editor of Statistics in Biosciences. She has served on a large number of committees of many statistical societies, and numerous NIH and NSF review panels.

Title: Scalable Integrative Analysis of Large-Scale Biobanks, Whole Genome Sequencing Studies and Functional Multi-Omic Data

Whole Genome/Exome Sequencing (WGS/WES) data and Electronic Health Records (EHRs), such as large scale national and institutional biobanks, have emerged rapidly worldwide. In this lecture, I will discuss the analytic tools and resources for scalable analysis of large scale biobanks and population-based Whole Genome Sequencing (WGS) association studies of common and rare variants by integrating WGS data with functional multi-omic data. I will also provide a demo of FAVOR (favor.genohub.org), a variant functional annotation online portal and resource that provides multi-faceted functional annotations of genome-wide 3 billion locations, and FAVORAnnotator, a tool that can be used to functionally annotate any WGS/WES studies. Cloud-based platforms for these resources will be discussed. The presentation will be illustrated using ongoing large scale whole genome sequencing studies and biobanks of quantitative, case-control, and time-to-event phenotypes, including the Genome Sequencing Program of the National Human Genome Research Institute and the Trans-Omics Precision Medicine Program from the National Heart, Lung and Blood Institute, and the UK Biobank and FinnGen.

Eminent Scholar Talk
Derrick Scott, Ph.D.
Sunday, August 7, 2022
4:15 - 4:35 PM
Logan

Bio

Dr. Scott is the Dean of the College of Natural and Health Sciences at Virginia State University after earning his B.S. from Virginia State University, his M.S. from Virginia Tech, and his Ph.D. from the University of South Carolina. He is passionate about creating opportunities for women and minorities and was recognized by Cell Mentor as one of the 1,000 Inspiring Black Scientists in America.

Dr. Scott's research involves bringing down the costs of expensive medicines by using informatics to identify target genes in Chinese hamster ovary (CHO) cell lines that will make the lines more stable and increase protein production. He has been leading major research, infrastructure, and workforce development initiatives with over $15M in grant support, most recently including the establishment and operation of the Delaware State University Molecular Diagnostic Laboratory where he served as the lab's Executive Director and assisted the university and surrounding communities stay safe via fast and frequent SARS-CoV-2 PCR testing.

Title: RNAseq: A Transcriptomics Approach to Lowering the Cost of Expensive Medicines.

CHO cells are the most important host cells used in manufacturing more than 75% of biologic medicines (also called biopharmaceuticals), with global sales over $120 billion per year. Nonetheless, a lack of understanding of the fundamental link between genome stability and the phenome significantly limits the ability of government, academic, and company laboratories to improve cell lines and ultimately product yields. We grew CHO cells under different industrially relevant growth conditions and used transcriptomics to identify genes that could potentially improve genome stability. This will have a direct effect to advance patient access to expensive medicines by helping to identify genes that can influence the protein level production of CHO cells.

Eminent Scholar Talk
Yuan Luo, Ph.D.
Monday, August 8, 2022
9:30-9:50 AM
Logan

Bio

Dr. Luo is currently Associate Professor at Department of Preventive Medicine, at Feinberg School of Medicine in Northwestern University. He is Chief AI Officer at Clinical and Translational Sciences Institute (NUCATS) and Institute for Augmented Intelligence in Medicine. Dr. Luo earned his PhD degree from MIT EECS with a math minor. He is a Fellow of American Medical Informatics Association (AMIA). He won the American Medical Informatics Association (AMIA) New Investigator Award in 2020.

Dr. Luo has been developing a novel suite of accurate, interpretable and generalizable models to integrate multi-modal health data (e.g., clinical and insurance claims data) for improving health care practice and advancing medical knowledge. He has been leading major research initiatives with >$10M in grant support, and has published over 100 peer-reviewed papers. His publications appear in leading journals including Nature Medicine, JAMA, AJRCCM, Circulation: Heart Failure, JAMIA, JBI etc. He has published in and/or served as PC members for top AI and informatics conferences including AAAI, KDD, CVPR, ACL etc. He has also been invited to give more than 50 keynotes and guest lectures at many top universities, think tanks, societies, industry labs.

Title: Machine Learning on Multi-Modal Healthcare Data

This talk will cover our recent progress on developing machine learning methods and applying them to multi-modal healthcare data with regional and national collaborative case examples. We will delve into different modalities of the healthcare data (e.g., unstructured clinical notes, structured EHR data, imaging data, genetic data etc.) and show how these data modalities can be individually and/or jointly mined to derive actionable intelligence. We will reflect on the lessons learned and argue for the need for developing flagship datasets to power high impact research. We will also argue for the need for a fresh perspective on AI/ML applications in healthcare and move from reactive to proactive machine learning.

Eminent Scholar Talk
Lana Garmire, Ph.D.
Monday, August 8, 2022
9:30-9:50 AM
Rittenhouse

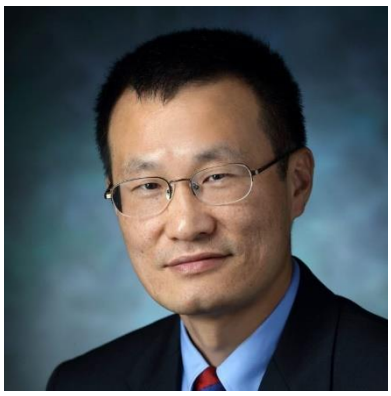
Bio

Lana Garmire is an associate professor of Computational Medicine and Bioinformatics from University of Michigan. Her research interest includes single-cell sequencing informatics and genomics; precision medicine through integration of multi-omics data types; novel modeling and computational methods for biomarker research, and population science. She has published over 90 papers in top quality journals and delivered over 70 invited talks. She is a standing member of BDMA study section and serves on the editorial advisory board for journals such as Genome Biology. Her work has been continuously funded by NIH in the last 10 years. Among the awards, noticeable ones include US Presidential Early Career Scientists and Engineers (PECASE), and fellow of American Institute of Medical and Biological Engineering (AIMBE).

Title: Pushing single cell data science towards clinical applications

Recently years we have seen a wealthy collection of computational methods developed to address challenges in single cell data science. However, an essential remaining question is how clinical domains can benefit from single cell research? I will debrief two research areas that we embarked to address the clinical need. First, I will describe a new drug recommendation method called ASGARD, which uses the patient scRNA-Seq data to repurpose drugs at the personalized level, exemplified by breast cancer, leukemia and COVID cases. Next, I will go over new discoveries on a large population cohort of single-cell images of breast cancer patients. We use tumor and tumor microenvironment information to reveal novel breast cancer subtypes with opposite prognosis outcomes to what previously had been generalized. Together these research projects highlight the promise of using single cell data science to guide personalized therapeutic treatment and to predict patient prognosis precisely.

Eminent Scholar Talk
Jiang Qian, Ph.D.
Monday, August 8, 2022
2:30-2:50 PM
Grand Ballroom



Bio

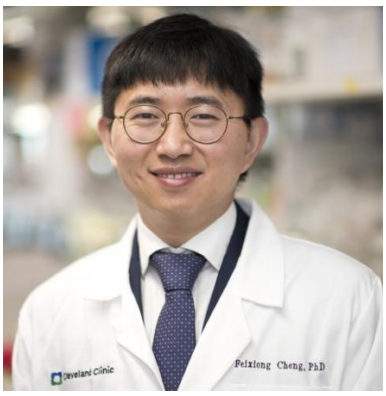
Jiang Qian, Ph.D., is the Karl H. Hagen professor of ophthalmology at the Wilmer Eye Institute. His research focuses on retinal gene regulation and the application of bioinformatics to the study of gene expression and regulation. Dr. Qian received an M.S. in computational biology from Shanghai Biochemistry Institute in Shanghai, China, and a Ph.D. in physical chemistry from the Max Planck Institute for Polymer Research in Mainz, Germany. He then completed a postdoctoral fellowship in bioinformatics at Yale University.

Title: Feedback loops in cell-cell communication

Intercellular communication (i.e. cell-cell communication) plays an essential role in multicellular organisms coordinating various biological processes. Previous studies discovered that feedback loops between two cell types are a widespread and vital signaling motif regulating development, regeneration, and cancer progression. While many computational methods have been developed to predict cell-cell communication based on gene expression datasets, these methods often predict one-directional ligand-receptor interactions from sender to receiver cells and are not suitable to identify feedback loops. Here we developed LRLoop, a new method for analyzing cell-cell communication based on bi-directional ligand-receptor interactions, where two pairs of ligand-receptor interactions are identified that are responsive to each other, and thereby form a closed feedback loop.

We first assessed LRLoop using bulk datasets and found our method significantly reduces the false positive rate seen with existing methods. Furthermore, we developed a new strategy to assess the performance of these methods in single-cell datasets and found that LRLoop produced a lower fraction of between-tissue interactions than traditional methods. Finally, we applied LRLoop to the single-cell datasets obtained from retinal development. We discovered many new bi-directional ligand-receptor interactions among individual cell types that potentially control proliferation, neurogenesis, and/or cell fate specification.

Eminent Scholar Talk
Feixiong Cheng, Ph.D.
Monday, August 8, 2022
2:30-2:50 PM
Logan



Bio

Dr. Feixiong Cheng is a systems pharmacologist and data scientist by training, with extensive expertise in analyzing, visualizing, and mining multimodal, high-dimensional heterogeneous data from real-world (e.g., electronic health records (EHRs) and health care claims) and experiments that profile the molecular state of human cells and tissues by genetics, genomics, transcriptomics (single-cell), proteomics, metabolomics, and interactomics (protein-protein interactions [PPIs] and chromatin interactions), for drug discovery and patient care with 15 years' experience. The primary goal of his lab is to combine tools from neurocomputation, artificial intelligence (AI), genetics and genomics (DNA/RNA sequencing), EHRs, network medicine, and experimental systems biology (PPIs) assays, to address the challenging questions toward understanding of human complex diseases (in particular for Alzheimer's disease), which could have a major impact in identifying novel real-world data-driven diagnostic biomarkers and therapeutic targets for precision medicine drug discovery and patient care (Two phase II trials have been initialized). Dr. Cheng has extensive experience in various aspects of large-scale human genome sequencing and multi-omics studies, including the Alzheimer's Disease Sequencing Project (ADSP), The Cancer Genome Atlas (TCGA), TopMed, PVDOMICS, and others. Dr. Cheng has created multiple multi-omics and EHR methodologies and successfully applied them for Alzheimer's disease drug discovery: (1) in silico network medicine-based discovery combined with insurance records data mining and patient iPSC-derived models identifies sildenafil (Viagra) as a candidate drug for Alzheimer's disease (Nature Aging 2021, PMID: 35572351) and a Phase II trial has been initialized; (2) multimodal single-cell/nucleus transcriptomics analysis combined with insurance records data mining identifies fluticasone and mometasone (approved asthma drugs) as candidate treatments for Alzheimer's disease (Genome Research 2021, PMID: 33627474); (3) insurance records data mining combined with mouse models identifies salsalate and diflunisal as candidate treatments for Alzheimer's disease and Traumatic brain injury (TBI) via reducing acetylated tau (Cell 2021, PMID: 33852912); and (4) AI-based multimodal analysis of genetic and genomic combined with EHR data identified pioglitazone (anti-diabetic drug) as candidate treatment for AD (Alzheimer's Research &

Therapy 2022, PMID: 35012639). He has served as keynote speakers or invited speakers in over 30 international and national conferences, including the FDA Scientific Computing Board 2020 and 2021 NIH Alzheimer's Research Summit: Path to Precision Medicine for Treatment and Prevention. He has received several Awards, including 2021 HHMI Gilliam Graduate Mentor Award (mentor graduate students under-represented background), 2020 ADDF Scholarships Winners, and NIH Pathway to Independence Award (K99/R00). In summary, his lab has established cutting-edge network medicine and systems pharmacology methodologies (Nature Biotechnology 2022, Nature Aging 2021, Nature Genetics 2021, PLOS Medicine 2021, Genome Biology 2021, PLOS Biology 2020, Lancet Digital Health 2020, Nature Communications 2019a, 2019b and 2018) to identify novel targets and repurposed drugs and combination therapies for a variety of human diseases, in particular for Alzheimer's disease.

Title: Harnessing Endophenotypes and Network Medicine for Discovery of Pathobiology and Drug Repurposing in Alzheimer's Disease

High-throughput DNA/RNA sequencing technologies have rapidly led to a robust body of genetic and genomic data in multiple national Alzheimer's disease (AD) genome projects, such as the Alzheimer's Disease Sequencing Project (ADSP) and the Alzheimer's Disease Neuroimaging Initiative (ADNI); however, the predisposition to AD involves a complex, polygenic, and pleiotropic genetic architecture. Recent advances in genetics and systems biology have showed that AD is governed by network-associated molecular determinants (termed disease module) of common endotypes or endophenotypes (e.g., Amyloid and Tau). Approaching AD with a simplistic single-target approach has been demonstrated effective for developing symptomatic therapies but ineffective when attempted for disease modification. Therapeutic approaches by specifically modulating genetic risk genes are essential for development of disease-modifying treatments in AD. However, existing data, including genomics, transcriptomics, proteomics, and interactomics (protein-protein interactions [PPIs]), have not yet been fully utilized and integrated to explore the roles of targeted therapeutic development for AD. Understanding AD from the point-of-view of how human interactome perturbations underlie the disease is the essence of network medicine. The main hypothesis of the AD network medicine is that cellular networks altered by genetic variants gradually rewire throughout disease pathogenesis and progression. Systematic identification and characterization of underlying AD pathogenesis and disease modules will serve as a foundation for identifying disease-modifying targets for AD. Integration of the genome, transcriptome, proteome, and the human interactome are essential for such identification. This seminar will introduce protein-protein interactome network-based, multimodal omics analysis technologies established by Cheng's lab to identify novel drug targets and repurpose existing drugs for Alzheimer's disease. Dr. Cheng will illustrate how his team combines tools from network medicine, endophenotype models, artificial intelligence (AI), multi-omics (GWAS, genomics, transcriptomics, and proteomics), and electronic health records (EHRs), to identify potential drug targets and repurposable drugs using Alzheimer's disease as a prototypical example.

Eminent Scholar Talk
Li Shen, Ph.D.
Tuesday, August 9, 2022
9:30-9:50 AM
Logan

Bio

Li Shen is a Professor of Informatics and the Interim Director of the Informatics Division in the Department of Biostatistics, Epidemiology and Informatics at the Perelman School of Medicine in the University of Pennsylvania. He obtained his Ph.D. degree in Computer Science from Dartmouth College. His research interests include medical image computing, biomedical and health informatics, machine learning, network science, imaging genomics, multi-omics and systems biology, Alzheimer's disease, and big data science in biomedicine. He has authored 300+ peer-reviewed articles in these fields. His work has been continuously supported by the NIH and NSF. His current research program is focused on developing and applying informatics, computing and data science methods for discovering actionable knowledge from complex biomedical and health data (e.g., genetics, omics, imaging, biomarker, outcome, EHR, health care), with applications to complex disorders such as Alzheimer's disease. He has served on a variety of scientific journal editorial boards, grant review committees, and organizing committees of professional meetings in medical image computing and biomedical informatics. He served as the Executive Director of the Medical Image Computing and Computer Assisted Intervention (MICCAI) Society between 2016 and 2019. He is a fellow of the American Institute for Medical and Biological Engineering.

Title: Integrating imaging and genomics data for gene discovery in Alzheimer's disease

Alzheimer's disease (AD) is a national priority, with 5.8 million Americans affected at an annual cost of $250+ billion and no available cure. Effective strategies are urgently needed to discover new AD genes for disease modeling and drug development. Studying AD genetics using multimodal imaging and genomics data is becoming a rapidly growing field with distinct advantages in power over categorical diagnosis under imaging and genomics traits as well as in capturing new insights into disease mechanism and heterogeneity from genetic determinants to molecular signatures, to brain imaging biomarkers, and to AD outcomes. In this talk, we will discuss statistical and informatics strategies for discovering AD risk and protective genes

through analyzing multidimensional genetics, genomics, imaging, and outcome data from landmark and local AD biobanks. We show that the wide availability of these rich biobank data, coupled with advances in biomedical statistics, informatics and computing, provides enormous opportunities to contribute significantly to gene discovery in AD and to impact the development of new diagnostic, therapeutic and preventative approaches.

Concurrent-Workshops
Sunday, August 7, 2022
1:00-2:45 PM
Grand Ballroom

Deep Learning: Methods and Biomedical Applications

Dr. Martin Renqiang Min, NEC-Labs

The analysis of tensor data, i.e., arrays with multiple directions, has become an active research topic in the era of big data. Datasets in the form of tensors arise from a wide range of applications, such as neuroimaging, genomics, and computational imaging. Tensor methods also provide unique perspectives to many high-dimensional problems, where the observations are not necessarily tensors. Problems with high-dimensional tensors generally possess distinct characteristics that pose unprecedented challenges to the data science community. There are strong demands to develop new methods to analyze the high-dimensional tensor data.

In this talk, we discuss how to perform SVD, a fundamental task in unsupervised learning, on general tensors or tensors with structural assumptions, e.g., sparsity, smoothness, and longitudinally. Through the developed frameworks, we can achieve accurate denoising for 4D scanning transmission electron microscopy images; in longitudinal microbiome studies, we can extract key components in the trajectories of bacterial abundance, identify representative bacterial taxa for these key trajectories, and group subjects based on the change of bacteria abundance over time. We also illustrate how we develop new statistically optimal methods and computationally efficient algorithms that exploit useful information from high-dimensional tensor data based on the modern theories of computation and non-convex optimization.

Concurrent-Workshops
Sunday, August 7, 2022
3:00-4:45 PM
Grand Ballroom

Multimodal data integration and analysis

Dr. Daifeng Wang, University of Wisconsin-Madison
Dr. Anru Zhang, Duke University

Modern healthcare is equipped with many scientific studies that produce large-scale biomedical data and render healthcare a data-driven service. Based on personal medical records, physiological data, and personal genomics data, precision healthcare provides personalized diagnosis, personalized medicine, and personalized optimal treatment strategies. In this talk, first I will introduce data-driven precision medicine, especially deep learning approaches to peptide-MHC interaction prediction and immunotherapy with T-cell receptor (TCR) engineering. Then I will talk about peptide-MHC interaction motif analysis and TCR optimization with deep reinforcement learning. Finally, I will conclude this talk by discussing some ongoing research directions of data-driven precision immunotherapy.

Concurrent-Workshops
Sunday, August 7, 2022
10:30 AM-4:45 PM
Logan

International Workshop on Translating Scientific Discoveries into Action (I2A)

In conjunction with the 10th International Conference on Intelligent Biology and Medicine (ICIBM2022),https://icibm2022.iaibm.org/

Introduction
Advancements in computation algorithms and biomedical technology have revolutionized biomedical research and generated massive data, knowledge, and scientific discoveries. However, the translation of the scientific knowledge to improve clinical practice, public health, drug development, and health policy has been limited by various factors in the complex healthcare environment. Researchers, practitioners, policymakers, pharmacists, industrial partners, and other stakeholders have been devoting a significant amount of time and effort to identifying the best approaches for transformative research and challenges from scientific discoveries to the bedside.
The 2022 International Workshop on Translating Scientific Discoveries into Action (TSD2A) will present cutting-edge informatics research and applications to clinical practice, public health, drug development, and health policymaking. It will provide close interactions among all the stakeholders to shape future application-oriented research into practice. It will provide a unique opportunity to share and apply transformative biomedical research outcomes to patient care and public health practices.
Workshop Organizing Committee:
Chair: Huanmei Wu, Professor and Department Chair, Assistant Dean for Global Engagement, Temple University College of Public Health
Publication Chair: Jay Patel, Assistant Professor, Temple University

10:30-11:00	**Networking**
Theme 1: Digital Twins for Health	
11:00 - 11:15	***The Emerging Digital Twins for Health: the technology and socio-ethical implications*** **Huanmei Wu, Temple University College of Public Health**
11:15 - 11:30	***Event labeling in digital twin for health*** **Hongfang Liu, Mayo Clinic**
11:30 - 11:45	***Patient-specific model for the metabolic panel of a cancer patient*** **Qi Wang, University of South Carolina**

11:45 - 12:00	***Digital Twins for Predictive Health: Vision and Roadmap*** **Jun Deng, Yale University School of Medicine**
12:00-1:00	***Lunch Break***
1:00 - 1:40	**Keynote Speech** ***Biomedical natural language processing: translating research to practice*** **Hua Xu, University of Texas - Houston, School of Biomedical Informatics**
Theme 2: Clinical Informatics and NLP	
1:40 - 1:55	***Discovering efficacy and safety of dietary supplements from multimodal data sources*** **Rui Zhang, University of Minnesota - Twin Cities**
1:55 - 2:10	***Application of informatics methods in dentistry to improve patient care and outcomes*** **Jay Patel, Temple University College of Public Health, Health Informatics**
2:10 - 2:25	***Enhance clinical data repository using natural language processing*** **Yonghui Wu, University of Florida**
2:25 - 3:40	***Zero-shot and Few-shot Learning to Address the Issue of the Lack of Labeled Data in Clinical Natural Language Processing*** **Yanshan Wang, University of Pittsburgh**
2:40 - 3:00	***Coffee Break***
Theme 3: Real-world Applications	
3:00 - 3:20	**ICIBM Eminent Scholar Presentation** ***RNAseq: A Transcriptomics Approach to Lowering the Cost of Expensive Medicines*** **Derrick Scott, Virginia State University**
3:20 - 3:35	***Evidence-based Law and Policymaking in Public Health*** **Jennifer Ibrahim, Temple University College of Public Health**
3:35 - 3:50	***Improving and Transforming Public Health Informatics through Community Engagement*** **Omar Martinez, University of Central Florida**
3:50 - 4:05	***How many roads must a man walk down, before translating biomedical discoveries into action?*** **Lixia Yao, Merck**
4:05 - 4:45	**Closing Keynote Speech** **What Problem Are We Trying to Solve? – Challenges and Strategies When Implementing Informatics in the Real World** **David Fleece, Temple University School of Medicine** **Chief Medical Information Officer, Temple University Hospital**
Dinner for Workshop Presenters (sponsored by Temple University College of Public Health)	

Concurrent-Workshops
Sunday, August 7, 2022
1:00-4:00 PM
Rittenhouse

Big data with statistical approaches

Dr. Yichuan Zhao, Georgia State University (Host)

Dr. Tian Tian, Children's Hospital of Philadelphia

Complex hierarchical structures in single-cell genomics data unveiled by deep hyperbolic manifold learning.

Delineating cell development is a critical analysis in single-cell genomic data. Numerous analytical methods have been developed; however, most are based on Euclidean space, which would distort the complex hierarchical structure of cell differentiation. Recently, methods acting on hyperbolic space have been proposed to visualize hierarchical structures in single-cell RNA-seq (scRNA-seq) data and proved to be superior to methods acting on Euclidean space. However, these methods have fundamental limitations and are not optimized for the large highly sparse single-cell count data. To address these limitations, we propose scDHMap, a model-based deep learning approach to visualize the complex hierarchical structures of scRNA-seq data in low dimensional hyperbolic space. The evaluations on extensive simulation and real experiments show that scDHMap outperforms existing dimensionality reduction methods in various common analytical tasks as needed for scRNA-seq data, including revealing trajectory branches, batch correction, and denoising highly dropout counts. In addition, we extend scDHMap to visualize single-cell ATAC-seq data.

Dr. Jung-Ying Tzeng, North Carolina State University

Association Methods for Biobank Studies: Scalable Gene-Environment Interaction (GxE) Tests and Copy Number Variant (CNV) Association Tests

Biobank comprises rich genetic and non-genetic information of large samples, and facilitates gene-environment interaction (GxE) studies and genetic association beyond single nucleotide polymorphisms. Here we present two recent work related to biobank studies. In the first work, we introduce SEAGLE to permit GxE variance component (VC) test for biobank-scale data, a widely used strategy to boost overall G×E signals from a genomic region. SEAGLE employs modern matrix computations to efficiently calculate the test statistic and p-value of the GxE VC test without imposing additional assumptions or relying on numerical approximations, and can accommodate

sample sizes in the order of 10^5. The second work focuses on CNV analysis, which requires special attentions because CNVs vary in dosage and length, have no natural “locus” definition, and can have heterogeneous etiological effects depending on the regions disrupted. We introduce CONCUR that treats CNVs of each individual as curve data over genomic locations and assesses association using a kernel machine framework. CONCUR evaluates CNV associations by comparing individuals' copy number profiles across genomic regions using the proposed "common area under the curve kernel", captures the effects of CNV dosage and length, and accommodates between- and within-position etiological heterogeneity without defining CNV loci. We illustrate the utility of the work using real data analyses on Taiwan Biobank data.

Dr. Xinlei (Sherry) Wang, Southern Methodist University

Bayesian Multiple Instance Classification Based on Hierarchical Probit Regression

In multiple instance learning (MIL), the response variable is predicted by features (or covariates) of one or more instances, which are collectively denoted as a bag. Learning the relationship between bags and instances is challenging because of the unknown and possibly complicated data generating mechanism regarding how instances contribute to the bag label. MIL has been applied to solve a variety of real-world problems, which have been mostly focused on supervised tasks, such as molecule activity prediction, protein binding affinities prediction, object detection, and computer-aided diagnosis. However, to date, the majority of the off-the-shelf MIL methods are developed in the computer science domain, and they focus on improving the prediction performance while spending little effort on explainability of the algorithm. We propose a Bayesian multi-instance learning model based on probit regression (MICProB), which contributes a significant portion to the suite of statistical methodologies for MIL. We evaluate the performance of MICProB against 15 benchmark methods and demonstrate its competitiveness in simulation and real data examples. In addition to its capability of identifying primary instances, as compared to existing optimization-based approaches, MICProB also enjoys great advantages in providing a transparent model structure, straightforward statistical inference of quantities related to model parameters, and favorable interpretability of covariate effects on the bag-level response.

Concurrent Session – Integrative multi-omics view of disease: from genetic variations to epigenetic dysregulation
Monday, August 8, 2022
9:30 AM - 12:25 PM
Grand Ballroom

Microbiome and tumor microenvironment crosstalk in the gastric cancer

Chao Zhang, Boston University School of Medicine

Gastric cancer carcinogenesis is associated with chronic inflammation, most commonly the result of Helicobacter pylori chronic infection in the stomach antrum. The development of gastric cancer in the context of chronic H. pylori infection is multifactorial, encompassing both bacterial factors and the altered immune microenvironment. However, a comprehensive analysis of the relation between inflammation and host microbial population in patient tissue samples has not previously been explored. We proposed an unbiased study to evaluate the relationships among microbiome composition, host immune response and genomic characterization from next generation sequencing of gastric biopsy samples. By investigating 152 collected samples and TCGA data, we discovered the bi-directional interaction between gastric microbiome and local immunity. We have not only identified a pro-inflammatory immune response signature associated with H. pylori infection, confirmed and validated by orthogonal assays. But also, the tumor immune microenvironment also could affect the local microbiome diversity.

Keyword: Cancer, Microbiome, TCGA, Whole genome sequencing, Inflammation

---

Specific role of CTCF in oncogenic transcriptional dysregulation

Chongzhi Zang, Associate Professor University of Virginia

Abstract:
Transcriptional dysregulation is a critical process in oncogenesis. Such functional changes in the 3-dimensional genome are context-specific and involve numerous nuclear protein factors. CCCTC-binding factor (CTCF) is a zinc finger protein that binds DNA and can induce DNA looping, functioning as a chromatin insulator by anchoring at topologically associating domain (TAD) boundaries and blocking cross-domain interactions. Disruption of individual CTCF binding in the genome can cause aberrant chromatin interaction and differential gene expression in many cell systems and cancer types. We systematically analyzed over 700 CTCF ChIP-seq profiles across human tissues and cancers, and identified cancer-specific CTCF binding patterns in 6 cancer types. We found that cancer-specific lost and gained CTCF binding events are associated largely with altered chromatin interactions, but only partially with DNA methylation changes and rarely with DNA sequence mutations. Instead, cancer-specific CTCF

binding events are likely induced by oncogenic transcription factor (TF) at super-enhancers, with a role to maintain phase-separated transcriptional condensates and to regulate oncogenic gene expression program. We argue that CTCF binding alteration can be used as a functional epigenetic signature of cancer.

Presentation Title: Linking genetic variants to kidney disease via the epigenome

Hongbo Liu[1,2,3]
Affiliation: [1]Department of Medicine, Renal Electrolyte and Hypertension Division, University of Pennsylvania, Philadelphia, PA, USA
[2]Institute of Diabetes Obesity and Metabolism, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA

Abstract:
More than 800 million people suffer from kidney disease, yet the mechanism of kidney dysfunction is poorly understood. Here we define the genetic association with kidney function in 1.5 million individuals and identify 878 (126 novel) loci. We map the genotype effect on the methylome in 443 kidneys, transcriptome in 686 samples, and single-cell open chromatin in 57,229 kidney cells. Heritability analysis reveals that methylation variation explains a larger fraction of heritability than gene expression. We present a multi-stage prioritization strategy, and prioritize target genes for 87% of kidney function loci. We highlight key roles of proximal tubules and metabolism in kidney function regulation. Furthermore, the causal role of SLC47A1 in kidney disease is defined in mice with genetic loss of Slc47a1 and in human individuals carrying loss-of-function variants. Our findings emphasize the key role of bulk and single-cell epigenomic information in translating genome-wide association studies into identifying causal genes, cellular origins and mechanisms of complex traits.

Keywords: GWAS, eQTL, meQTL, DNA methylation, Open chromatin, Chronic kidney disease

---

RNA m6A landscape reveals strong regulatability of tumor suppressor expression

Kaifu Chen[1,2,]
[1]Department of Pediatrics, Harvard Medical School, Department of Cardiology
[2]Boston Children's Hospital

Abstract:
Cancer genes were known to display unique epigenetic features on chromatin of benign cells. Investigations into these features are making it increasingly clear that cancer genes differ from other genes regarding the mechanisms regulating their transcription. It is yet unknown whether cancer genes have a unique epitranscriptomic feature on RNAs and thus differ from other genes in post-transcriptional regulation of their RNA expression. Here we found RNAs of tumor suppressor genes tended to decay fast in multiple benign cell types when compared with other RNAs. Consistent with a negative effect of m6A modification on RNA stability, we observed preferential deposition of m6A on tumor suppressor RNAs. With frequent transcription, the fast RNA decay of tumor suppressors did not lead to low expression in benign cells.

However, abundant m6A and fast decay of tumor suppressor RNAs both tended to be further enhanced in prostate cancer cells relative to benign prostate epithelial cells. This enhancement correlated with a down regulation of tumor suppressor expression. Further, knockdown of m6A methyltransferase METTL3 and reader protein YTHDF2 in prostate cancer cells posed stronger effect on tumor suppressor RNAs than on other RNAs. These results indicated a strong expression maneuverability of tumor suppressors mediated by abundant m6A modification on RNAs.

Keywords: RNA modification, m6A, histone Modification, cancer, epigenomics, epitranscriptomics.

Harnessing big data to characterize toxicity of immunotherapy

Ying Jing[1], Yuan Liu[2], Jingwen Yang[2], Leng Han[1,2,3]

[1]Department of Biochemistry and Molecular Biology, The University of Texas Health Science Center at Houston McGovern Medical School, Houston, TX, USA
[2]Center for Epigenetics and Disease Prevention, Institute of Biosciences and Technology, Texas A&M University, Houston, TX, USA.
[3]Department of Translational Medical Sciences, College of Medicine, Texas A&M University, Houston, TX, USA.

Abstract
Immune-checkpoint inhibitors (ICIs) have transformed patient care in oncology but are associated with a unique spectrum of organ-specific inflammatory toxicities known as immunerelated adverse events (irAEs). Given the expanding use of ICIs, an increasing number of patients with cancer experience irAEs, including severe irAEs. Proper diagnosis and management of irAEs is important to optimize the quality of life and long-term outcomes of patients receiving ICIs; however, owing to the substantial heterogeneity within irAEs, and despite multicentre initiatives, performing clinical studies with a sufficient cohort size is challenging. Very recently, we utilized a strategy that to combine the power of real-world data and omics data and evaluated associations between multi-omics factors and irAE reporting odds ratio across different cancer types. We identified a bivariate regression model of LCP1 and ADPGK that can accurately predict irAE, and further validated LCP1 and ADPGK as biomarkers in an independent patient-level cohort[1]. Utilizing similar strategy, we also demonstrated the associations of irAEs and antibiotic during anti-PD-1/PD-L1 therapy across a wide spectrum of cancers by analyzing multi-source data, suggesting that administration of antibiotics should be carefully evaluated in cancer patients treated by anti-PD-1/PD-L1 to avoid potentially increasing irAE risk[2]. Furthermore, we exploited large-scale single-cell analysis to report potential on-target, off-tumor toxicity landscape for chimeric antigen receptor (CAR) targets across a wide range of tissues[3]. Taken together, our studies demonstrated that aggregate clinical data, real-world data (such as data on pharmacovigilance or from electronic health records) and multi-omics data are alternative tools well suited to investigating the underlying mechanisms and clinical presentations of toxicities[4].

Related Publications:
1. Jing et al. Multi-omics prediction of immune-related adverse events during checkpoint

immunotherapy. **Nature Communications**, 2020, 11: 4946, PMID: 33009409
2. Jing et al. Association of antibiotic treatment with immune-related adverse events in patients with cancer receiving immunotherapy. **Journal of Immunotherapy of Cancer,** 2022, 10(1):e003779. PMID: 35058327
3. Jing et al. Expression of chimeric antigen receptor therapy targets detected by single-cell sequencing of normal cells may contribute to off-tumor toxicity. Cancer Cell, 2021. PMID: 34678153
4. Jing et al. Harnessing big data to characterize immune-related adverse events. Nature Reviews Clinical Oncology, 2022. PMID: 350396791

Keywords: Immunotherapy; Immune-checkpoint inhibitors; chimeric antigen receptor; big data; toxicity

---

Cobind: quantify the magnitude of genomic collocation

Tao Ma[1], Lingyun Guo[2], Liguo Wang[1,3],*

[1] Division of Computational Biology, Mayo Clinic College of Medicine and Science, Rochester, MN 55905, USA;
[2] Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA;
[3] Bioinformatics and Computational Biology Graduate Program, University of Minnesota Rochester, Rochester, MN 55904, USA

Collocated genomic intervals indicate biological association. The commonly used approach to evaluate the strength of collocation generally applies an arbitrary threshold to decide the amount (or proportion) of overlapped genomic regions. Such thresholded approach ignores the size of each genomic interval, leads to inaccurate, biased, non-reproducible, and incomparable results. Here we developed the cobind package, which provide six threshold-free metrics to rigorously measure the magnitude of genomic collocation. When applied these new approaches to genomic intervals identified from transcription factor chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-seq), we demonstrated that the collocation coefficient (C) and normalized pointwise mutual information (NPMI) as the best measurements to quantify genomic collocations, and these new approaches successfully nominated all the cohesin proteins –CTCF's experimentally validated co-factors–as the top five transcription factors that col-localized with CTCF bindings. When further applied to cis-regulatory regions identified from bulk and single-cell Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), our methods could effectively nominate known and novel master regulators from prostate cancer and oligodendrocyte cells.

Keywords: genomic interval, overlapping, collocation, ChIP-seq, ATAC-seq, master regulator.

* Corresponding Author: Liguo Wang, Wang.Liguo@mayo.edu

Dissect human diseases with high throughput single cell nanopore sequencing

Ruli Gao, Department of Biochemistry and Molecular Genetics, Northwestern University, Chicago, IL, USA 60611

Human tissues represent complex ecological systems of diverse cell types with dynamic genetic evolution (cell lineages) and transcriptional remodeling (cell fates). However, there is a lack of robust methods for measuring both genotypes and phenotypes of same cells to precisely trace cellular dynamics during disease development. We developed a high throughput single cell nanopore sequencing method, HT-scNanoSeq to detect full length cDNAs for thousands of single cells. In concert with this technology, we developed a computational tool, scNanoGPS to perform independent barcode assignment without paralleled short reads curation and calculate transcriptional activities and genetic alterations simultaneously from same cells. Our methods detected cell type and cell state specific mutations, splicing isoforms as well as transcriptional programs, which enabled direct mapping of cell fate transitions onto cell lineages in human disease development.

Extend the health-span: decipher the epigenome of aging and leukemogenesis

Sheng Li, The Jackson Laboratory for Genomic Medicine

Acute myeloid leukemia (AML) is a deadly blood cancer that occurs mainly in adults over 65 and is associated with the abnormal overproliferation of hematopoietic stem cells, a common age-related condition called 'clonal hematopoiesis' (CH). Although we know that aging and specific gene mutations in genes encoding epigenetic regulators in the hematopoietic stem cells contribute to the development of CH and its progression to AML, we do not understand how old age and CH interact to promote the evolution of AML. As the US population ages, there is an urgent unmet need for new therapeutic strategies to mitigate CH evolution to leukemia. We developed multiple bioinformatics tools to quantify the epiallele repertoire shift in somatic evolution and infer cell-to-cell epigenetic heterogeneity by extracting the single-molecule, single-base resolution DNA methylation patterns from bisulfite sequencing data. Using our tools, we showed that hematopoietic clones from relapsed AML patients manifest a selection of malignant cells harboring specific epigenetic alleles, and such alleles are associated with unfavorable outcomes. Moreover, AML epigenetic heterogeneity provides pre-malignant HSPC with an additional layer of fitness beyond genetic heterogeneity before the malignant transformation. Lastly, the hematopoietic stem cells from aged mice exhibit significantly higher epigenetic and transcriptomic heterogeneity by single-cell transcriptome and chromatin accessibility data analysis. These results collectively suggest a putative epigenetic mechanism for aging and TET2 mutation to jointly impact the evolvability of mutant HSC and thus contribute to clonal expansion and leukemogenesis via enhanced epigenetic heterogeneity, laying the foundation for developing 'evolution-blocking' strategies to prevent leukemia in high-risk aging populations.

Keywords: Bioinformatics, epigenetic heterogeneity, leukemogenesis, cancer evolution, hematopoietic aging.

Concurrent Session – Machine learning in Cancer
Monday, August 8, 2022
9:30 - 11:45 AM
Logan

Risk Stratification for Breast Cancer Patient by Simultaneous Learning of Molecular Subtype and Survival Outcome Using Genetic Algorithm-Based Gene Set Selection

Bonil Koo[1], Dohoon Lee[2], Sangseon Lee[3], Inyoung Sung[1] and Sun Kim [1,4,5,6,*]
[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[2]Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea
[3]Institute of Computer Technology, Seoul National University, Seoul, Republic of Korea
[4]Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
[5]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea
[6]AIGENDRUG Co., Ltd., Seoul, Republic of Korea
*Correspondence: sunkim.bioinfo@snu.ac.kr
[†]This paper is an extended version of our paper published in ICIBM 2022.

Abstract
Patient stratification is a clinically important task because it allows us to establish and develop efficient treatment strategies for particular groups of patients. Molecular subtypes have been successfully defined using transcriptomic profiles and they are used effectively in clinical practice, e.g., PAM50 subtypes of breast cancer. Survival prediction contributed to understanding diseases and also identifying genes related to prognosis. It is desirable to stratify patients considering these two aspects simultaneously. However, there are no methods for patient stratification that consider molecular subtypes and survival outcomes at once. Here, we propose a methodology to deal with the problem. Genetic algorithm is used to select a gene set from transcriptome data, and their expression quantities are utilized to assign a risk score to each patient. The patients are ordered and stratified according to the score. A gene set was selected by our method on a breast cancer cohort (TCGA-BRCA), and we examined its clinical utility using an independent cohort (SCAN-B). In this experiment, our method was successful in stratifying patients with respect to both molecular subtype and survival outcome. We demonstrated that the orders of patients were consistent across repeated experiments, and prognostic genes were successfully nominated. Additionally, it was observed that the risk score can be used to evaluate the molecular aggressiveness of individual patients.

Keywords: patient stratification; molecular subtype; survival outcome; genetic algorithm; gene set selection

A Novel Bayesian Framework Infers Driver Activation States and Reveals Pathway-oriented Molecular Subtypes in Head and Neck Cancer

Zhengping Liu[1,2], Chunhui Cai[1*], Xiaojun Ma[1], Jinling Liu[3,4], Lujia Chen[1], Vivian Wai Yan Lui[5], Gregory F. Cooper[1,6], Xinghua Lu[1,6]

[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, PA, United States of America
[2]School of Medicine, Tsinghua University, Beijing, China
[3]Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, MO, United States of America
[4]Department of Biological Science, Missouri University of Science and Technology, MO, United States of America
[5]Georgia Cancer Center, and Department of Medicine, Medical College of Georgia, Augusta University, GA, United States of America
[6]UPMC Hillman Cancer Center, University of Pittsburgh Medical Center, Pittsburgh, PA, United States of America

* Corresponding Author
Email: chunhuic@pitt.edu

Abstract
Head and neck squamous cell cancer (HNSCC) is an aggressive cancer resulting from heterogeneous causes. To reveal underlying drivers and signaling mechanisms of different HNSCC tumors, we developed a novel Bayesian framework to identify drivers of individual tumors and infer the states of driver proteins in cellular signaling system in HNSCC tumors. First we systematically identify causal relationships between somatic genome alterations (SGAs) and differentially expressed genes (DEGs) for each HNSCC tumor using the tumor-specific causal inference (TCI) model. Then, we developed machine learning models that combine genomic and transcriptomic data to infer the functional activation states of the proteins of driver genes in tumors, which enable us to represent a tumor in the space of cellular signaling systems. We have discovered 4 mechanism--oriented subtypes of HNSCC, which show distinguished patterns of activation state of HNSCC driver proteins, and importantly, this subtyping is orthogonal to previously reported transcriptomic-based molecular subtyping of HNSCC. Further, our analysis revealed driver proteins that are likely involved in oncogenic processes induced by HPV infection, even though they are not perturbed by genomic alterations in HPV+ tumors.

Robust personalized classifier improves the prediction of breast cancer metastasis

Nahim Adnan[1] and Jianhua Ruan[*]

[1]Department of Computer Science, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA.

*Correspondence: jianhua.ruan@utsa.edu

Department of Computer Science, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249,USA

Abstract
Accurate detection of breast cancer metastasis in early stage of diagnosis of cancer is very crucial to reduce death in breast cancer in women. With the availability of gene expression dataset, many machine learning models have been proposed to detect metastasis using thousands of genes simultaneously. However, the prediction accuracy of the models using gene expression often suffer from the diverse molecular characteristics across different datasets. Additionally, breast cancer is known to have many subtypes which also hinders the performance of the models aimed at prediction for all subtypes. To overcome the heterogeneity nature of breast cancer, we propose personalized classifier which is trained on a subset of most similar and dissimilar patients from the training dataset for predicting a specific patient from the test dataset. Results showed that our proposed approach significantly improved prediction accuracy compared to the models trained on the complete training dataset as well as models trained on specific subtypes. Our results also showed that personalized classifiers trained on both positively and negatively correlated patients outperformed classifiers trained only on positively correlated patients, which highlights the importance of selecting proper patient subsets for constructing personalized classifiers. Additionally, our proposed approach obtained more robust features compared to the other models, and was able to identify different features for different patients, making it a promising tool for designing personalized medicine for cancer patients.

Keywords: breast cancer; metastasis; personalized classifier

High Resolution Cell Type Deconvolution Reveals Cell Type Specific Molecular Mechanism of Cancer Radioresistance

Xiao Sun[1,2,3], Min Zhu[2,3 *], Xiaoyi Fei[2,3,4], Xueling Li[2,3 *]

[1]School of Electronic and Information Engineering, Anhui Jianzhu University, South Campus: No. 292 Ziyun Road, Shushan District, Hefei 230009, People's Republic of China.
[2]Anhui Province Key Laboratory of Medical Physics and Technology, Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei 230031, People's Republic of China.
[3]Hefei Cancer Hospital, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei 230031, People's Republic of China.
[4]School of Biomedical Engineering, Anhui Medical University, 81 Meishan Road, Shushan District, Hefei 230009, People's Republic of China.

*Correspondence: Min Zhu: mizhu@VIP.163.com and Xueling Li: xlli@cmpt.ac.cn. This work was performed in Institute of Health and Medical Technology, Hefei Institutes of Physical Science.

Abstracts
Tumor immune microenvironment and the heterogeneity of its cell type composition play important roles in cancer progression and radiotherapy treatment response and prognosis. However, how the

constitutes and ecosystem of the cancer immune microenvironment are associated with the outcomes of the radiotherapy are still not fully elucidated. The current single cell RNA sequencing is powerful for characterizing cellular heterogeneity. However, analyses of large sample cohorts are not yet practical with limited sample size due to its high expenses. Considering it, we integrated the single cell sequencing data and high resolution transcriptomics cell type deconvolution of rectal cancer before adjuvant radiotherapy for comparative analysis of the responders vs non-responders. Gene expression purification on 300 top bulk differentially expressed genes between resistant and sensitive patients to radiotherapy and 2659 signature genes from single cell sequencing data was performed with high resolution mode of CIBERSORTx. We obtained 216, and 217 purified cell type specific DEGs including 26 and 33 DEGs encoding receptor or ligand, respectively. The obtained cell type specific DEGs may inform molecular mechanism and receptor-ligand interaction mediated cell-cell communication through autocrine and paracrine. We validated our findings with literature search, which demonstrated the biomedical significance of our inferred cell type specific ligand-receptor interaction medicated cell-cell communication in the prediction of the radiotherapy. Our results can be further validated through analysis of rectal cancer spatial transcriptomics.

Modeling the relationship between gene expression and mutational signature

Limin Jiang[1], Hui Yu[1], Yan Guo[1*]

[1]Department of Internal Medicine, Comprehensive Cancer Center, University of New Mexico Albuquerque, NM, 87109, USA
*Corresponding Author: Yan Guo
Email: yanguo1978@gmail.com

Abstract
Mutational signatures computed from somatic mutations, allow an in-depth understanding of tumorigenesis and may illuminate early prevention strategies. Many studies have shown the regulation effects between somatic mutation and gene expression dysregulation. Thus, we hypothesized that there are potential associations between mutational signature and gene expression. We capitalized upon RNA-seq data to model 49 established mutational signatures in 33 cancer types. Both accuracy and area under the curve were used as performance measures in five-fold cross-validation. As a result, 475 models using unconstrained genes, and 112 models using protein-coding genes were selected for future inference purposes. An independent gene expression dataset on lung cancer smoking status was used for validation which achieved over 80% for both accuracy and area under the curve. These results demonstrate that the associations between gene expression and somatic mutations can translate into the associations between gene expression and mutational signatures.

Keywords: Mutational Signature; gene expression; support vector machine; random forest; extreme gradient boost

---

Concurrent Session – Informatics in team science: to lead, support, and educate
Monday, August 8, 2022
9:30 – 11:45 AM
Rittenhouse

The WVU Bioinformatics Core: Empowering Biomedical Research through Team Science in 'Almost Heaven' West Virginia

Gangqing (Michael) Hu, School of Medicine, West Virginia University

Abstract: The vision of the West Virginia University Bioinformatics Core is to advance biomedical research at WVU and beyond through team science. We offer timely and high-quality services on various sequencing data, primarily through collaborative research. The Core works with the WVU Genomics Core and the Marshall University Genomics and Bioinformatics Core to facilitate a one-stop, statewide service for library preparation, sequencing, and data analysis. During the past three years, the Core has addressed the challenge of staff shortages by involving graduate students in team science, by developing and offering a new graduate level bioinformatics class, and by referring some requests to external partners. A unique feature of the Core is that the director also maintains an independent research lab which has served to populate the applications of many state-of-the-art epigenetic assays across the campus. Current challenges will be to address an exponential growth in demand and to develop a sustainable mechanism for cost recovery and staff retention.

Team science in a large medical institution: the tradition of "the needs of the patient come first"

Zhifu Sun, Department of Quantitative Health Sciences, Mayo Clinic Rochester

Team science is critical in modern biomedical research and discovery or any other areas; however, it is easier said than done in practice when a large group of people comes from different expertise and speaks different "languages". At Mayo Clinic, the spirit of "the needs of the patient come first" works like "super glue" to keep everyone involved working together to achieve that ultimate goal. In this talk, I will highlight the excellent research environment to conduct team science at Mayo Clinic, share some good stories/projects, and also some challenges in a big organization.

NGS-related bioinformatics in Academy and Industry

Yaping Feng, Senior, Admera Health

The applications of NGS have been expanded into many fields such as: human healthcare, animal health, plant breeding and diversity, microbiome etc. The human healthcare applications include early diagnosis, biomarker discovery, personal medicine, prognosis, and vaccine/antibody development etc. The

corresponding bioinformatics tools and analysis came into being. The top three categories of the bioinformatics analysis provided by our company are RNAseq, customized gene panel, and WES, which account for ~80% of total. A closer look at these categories in the aspect of the industry/academy share shows that the industry-sourced bioinformatics analysis accounts for 62%, 91%, and 70% respectively. All other categories such as miRNA, Chip-seq, ATACseq, cut & run, 10x single cell, metagenome etc. are mostly from academic institutes. This limited view from our company indicates the bioinformatics analysis needs in the translational utilities of the top three categories are maturing and the analysis for other NGS methods are still mostly in academic usage or the early translational stage.

Education in bioinformatics: "how to position and balance?"

Jingwen Yan, School of Informatics and Computing, Indiana University–Purdue University Indianapolis

Biomedical informatics, built on top of the multi-scale multi-omic and healthcare data, is a highly interdisciplinary field drawing from biology, medicine, computer science, engineering, and beyond. Although the escalating volume of biomedical data is driving an unprecedented need in information management, decision support and advanced analytics, educational pathways into this field remain largely unclear given the challenges presented by its broad foundation. While it is impossible to learn everything of all related disciplines, education plans are expected to help students position their future career and balance their interdisciplinary skills toward that. The Department of BioHealth Informatics at IUPUI draws on expertise from disciplines such as biology, mathematics, statistics, and medicine to enhance our understanding of life sciences and advance health care. Students will study and pursue the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, driven by efforts to improve human health. Our program aims to 1) increase the awareness of educational opportunities and career paths in biomedical informatics, 2) cultivate the interest of students in learning biomedical informatics, and 3) build their confidence in staying in this field.

Biomedical informatics training in diverse environments

Li Liu, School of Health Solutions, Arizona State University

Arizona State University (ASU) is among the largest universities in the US with quickly expanding academic programs and student populations. The demands for biomedical informatics support come from multiple levels and continue to grow. However, the capital, human, and computational resources available to individual labs are limited. In this talk, I will share my experience of responding to opportunities and challenges in a public research university environment, focusing on involving students with diverse background in team science. With proper training, undergraduate and graduate students from different programs participate in projects ranging from core services to research studies, and from university-wide initiatives to community outreach. I will discuss the importance of fostering mutual respect and establishing leadership role of biomedical informatics researchers. By aligning the interest and skills of trainees and the needs of projects, we aim to improve resource allocation, student success, and research productivity.

Concurrent Session – Cancer informatics
Monday, August 8, 2022
2:30 - 4:40 PM
Grand Ballroom

CellCallEXT: analysis of ligand–receptor and transcription factor activities in cell–cell communications of tumor immune microenvironment

Shouguo Gao, Xingmin Feng, Zhijie Wu, Sachiko Kajigaya, Neal S Young

Hematopoiesis and Bone Marrow Failure Laboratory, Hematology Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, United States of America

§Corresponding author
Email addresses: SG: shouguo.gao@nih.gov

Abstract
Background: Single-cell RNA sequencing (scRNA-seq) data is useful for the decoding of cell-cell communications. CellCall is a tool to infer inte- and intracellular communication pathways by integrating paired ligand-receptor and transcription factor (TF) activities from steady state data, thus can not handle two-condition comparisons directly. For tumor and healthy status, it can only analzye cells from tumor and healthy controls individually and examine the ligand-receptor pairs only identified in tumor or healthy control. Further CellCall is greatly affected by gene expression specificity in tissues.
Results: CellCallEXT is an extension of CellCall that allows deciphering intercellular communication alteration and related internal regulatory signals based on scRNA-seq. Information in Reactome was retrieved and integrated with prior knowledge of ligand-receptor-TF signaling and gene regulation datasets of CellCall. It is designed to directly identify the L-R interactions that alter the expression profiles of downstream genes between two conditions, such as tumor and healthy controls. CellCallEXT was successfully applied to examine tumor and immune cell microenvironments and to identify the altered ligand-receptor pairs and downstream gene regulatory networks among immune cells. Application of CellCallEXT to scRNA-seq data from patients with deficiency of adenosine deaminase 2 also demonstrated its ability to impute dysfunctional intercellular communications and related transcriptional factor activities.

Availability: The tool and sample script are available on https://github.com/shouguog/cellcallEXT.
Contact: shouguo.gao@nih.gov

Identification of immuno-targeted combination therapies using explanatory subgroup discovery for cancer patients with EGFR wild-type (WT) gene

Olha Kholod[1], William Basket1, Danlu Liu[2], Jonathan Mitchem[1,3,4], Jussuf Kaifi[1,3], Laura Dooley[5] and Chi-Ren Shyu[1,2,*]

[1]MU Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65212, USA; okbm4@mail.missouri.edu (O.K); wibpp9@mail.missouri.edu (W.B.); shyuc@missouri.edu (CR.S.)
[2]Department of Electrical Engineering & Computer Science, University of Missouri, Columbia, MO 65212, USA; dltb9@mail.missouri.edu (D.L.)
[3]Department of Surgery, School of Medicine, University of Missouri, Columbia, MO 65212, USA; mitchemj@health.missouri.edu (J.M.), kaifij@health.missouri.edu (J.K.)
[4]Harry S. Truman Memorial Veterans' Hospital, Columbia, MO 65201, USA;
[5]Department of Otolaryngology, School of Medicine, University of Missouri, Columbia, MO 65212, USA; DooleyL@health.missouri.edu (L.D)

* Correspondence: shyuc@missouri.edu

Abstract
Background: phenotypic and genotypic heterogeneity are characteristic features of cancer patients. To tackle patients' heterogeneity, immune checkpoint inhibitors (ICIs) represent one of the most promising therapeutic approaches. However, approximately 50% of cancer patients that are eligible for treatment with ICIs will not respond well, especially patients with no targetable mutations. Over the years, multiple patient stratification techniques have been developed to identify homogenous patient subgroups, although, matching patient subgroup to treatment option
that can improve patients' health outcome remains a challenging task.
Methods: we extend our subgroup discovery algorithm to identify patient subpopulations that can potentially benefit from immuno-targeted combination therapies in four cancer types: Head and Neck Squamous Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), Lung Squamous
Carcinoma (LUSC) and Skin Cutaneous Melanoma (SKCM). We employ the proportional odds model to identify significant drug targets and corresponding compounds that increase the likelihood of stable disease versus progressive disease in cancer patients with EGFR wild-type (WT) gene.
Results: our pipeline identifies six significant drug targets and thirteen specific compounds for cancer patients with EGFR WT gene. Three out of six drug targets – FCGR2B, IGF1R and KIT – substantially increase the odds of having a stable disease versus progressive disease. Progression free survival (PFS) more than 6 months was a common feature among investigated subgroups.
Conclusions: our approach can help to better select responders for immuno-targeted combination therapies and improve health outcome for cancer patients with no targetable mutations.

Keywords: immuno-targeted combination therapies; subgroup discovery; cancer

---

Spatial transcriptomic analysis reveals associations between genes and cellular topology in breast and prostate cancers

Lujain Alsaleh[1], Chen Li[2], Justin L. Couetil[3], Kun Huang[1,3,5], Jie Zhang[3,5], Chao Chen[2], Travis S. Johnson[1,5,6,*]

[1]Department of Biostatistics and Health Data Science, Indiana University
[2]Department of Biomedical Informatics, Stony Brook University
[3]Department of Medical and Molecular Genetics, Indiana University
[4]Regenstrief Institute
[5]Melvin and Bren Simon Comprehensive Cancer Center, Indiana University
[6]Indiana Biosciences Research Institute

* Corresponding author is: Travis S. Johnson (johnstrs@iu.edu)

Abstract
Background: Cancer is the leading cause of death worldwide and one of the most studied topics in public health. Breast and prostate cancer are the most common cancers among women and men respectively. Gene expression and image features are independently prognostic of patient survival; but until the advent of spatial transcriptomics (ST), it was not possible to determine how gene expression of cells is tied to their spatial relationships (i.e., topology). Topology and its modern development, persistent homology, is a mathematical theory to quantify structures of objects arising in complex systems (e.g., cells in our problem). Although we have observed predictive power, features derived from topology are often not directly interpretable to a pathologist. Herein, we use integrative bioinformatics analysis techniques to correlate cell topology with molecular profiles and identify sets of topological features that represents histological components of the tumor and microenvironment (TME). We identify topology- associated genes (TAGs) that correlate with 700 image topological features (ITFs), in breast and prostate cancer 10x Visium spatial transcriptomics samples.
Method: Genes and image topological features are independently clustered using pheatmap package and correlated with each other using Pearson correlation matrix. This correlation matrix is visualized with R package pheatmap to show relationships between the two sets. Themes among genes correlated with ITFs are investigated by functional enrichment analysis using Toppgene, which are the most significant genes. We also use rank of correlation coefficients to identify similarities and differences in gene-topology correlations across breast cancer ST and in prostate cancer ST slides as well.
Result: TAGs corresponding to extracellular matrix (ECM) and Collagen Type I Trimer gene ontology terms common to both prostate and breast cancer. In breast cancer ST slides, we found ZAG-PIP Complex gene is correlated to breast cancer. IgA immunoglobulin complex found to have correlations with prostate cancer in prostate cancer ST slides.
Conclusion: We identified topology-associated genes in every ST slide regardless of cancer type. These TAGs are enriched for ontology terms, illustrating the biological relevance to our image topology features and their potential utility in diagnostic and prognostic models.

Clone phylogenetics reveals metastatic tumor migrations, maps, and models

Antonia Chroni[1,2,+], Sayaka Miura[1,2,+], Lauren Hamilton[1,2], Tracy Vu[1,2], Stephen Gaffney[4], Vivian Aly[1,2], Sajjad Karim[3], Maxwell Sanderford[1,2], Jeffrey P. Townsend[4,5,6], and Sudhir Kumar[1,2,3,*]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA.
[2]Department of Biology, Temple University, Philadelphia, PA.

[3]Center for Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia
[4]Department of Biostatistics, Yale University, New Haven, CT.
[5]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT.
[6]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT.

*Corresponding author:
Sudhir Kumar (s.kumar@temple.edu)

+Co-first authors:
Antonia Chroni (antonia.chroni@temple.edu) and Sayaka Miura (tuf78332@temple.edu)

Co-authors
Jeffrey Townsend (jeffrey.townsend@yale.edu), Lauren Hamilton (tuh03080@temple.edu), Tracy Vu (tug86468@temple.edu), Stephen Gaffney (stephen.gaffney@yale.edu), Vivian Aly (vivian.aly@temple.edu), Sajjad Karim (skarim1@kau.edu.sa), and Maxwell Sanderford (m.sanderford@temple.edu)

Abstract
Dispersal routes of metastatic cells are not medically detected or even visible until the later stages of cancer progression. Molecular evolutionary analysis of tumor variation provides a way to retrospectively infer metastatic migration histories and answer questions such as whether the majority of metastases are seeded by clones from primary tumors instead of pre-existing metastases and whether the evolution of metastases is generally consistent with any proposed models. We seek answers to these fundamental questions through a systematic patient-centric retrospective analysis that maps the dynamic evolutionary history of tumor cell migrations in many cancers. We analyzed tumor genetic heterogeneity present in 51 cancer patients and found that most metastatic migration histories are described by a hybrid of multiple models of metastatic tumor evolution. Synthesizing across metastatic migration histories, we found new tumor seedings by clones of pre-existing metastases as often as clones from primary tumors. There were also many clone exchanges between source and recipient tumors. Therefore, molecular phylogenetic analysis of tumor variation provides a retrospective glimpse into general patterns of metastatic migration histories in cancer patients.

Keywords: Tumor evolution; metastasis; molecular evolution; phylogenetics; phylodynamics; cancer

---

Genomic variants disrupt miRNA-mRNA regulation

Ellie Xi[1†], Judy Bai[1†], Klaira Zhang[1], Hui Yu[1], Yan Guo[1*]

[1]Department of Internal Medicine, University of New Mexico, Albuquerque NM 87131 †equal contribution

*corresponding author

Abstract
Micro RNA (miRNA) and its regulatory effect on messenger RNA (mRNA) gene expression are a major focus in cancer research. Disruption in the normal miRNA-mRNA regulation network can result in serious cascading biological repercussions. In this study, we curated miRNA-related variants from major genomic consortiums and thoroughly evaluated how these variants could exert their effects by cross-validating with independent functional knowledge bases. Nearly all known variants (more than 664 million) categorized by type (germline, somatic, epigenetic) were mapped to the genomic regions involved in miRNA-mRNA binding (miRNA seeds and miRNA-mRNA 3'-UTR binding sequence). Subsets of miRNA-related variants supported by additional functional evidence, such as expression Quantitative Trait Loci (eQTL) and Genome-Wide Association Study (GWAS), were identified and scrutinized. Our results show that variants in miRNA seeds can substantially alter the composition of a miRNA's target mRNA set. Various functional analyses converged to reveal a post-transcriptional complex regulatory network where miRNA, eQTL, and RNA-binding protein intertwined to disseminate the impact of genomic variants. These results may potentially explain how certain variants affect disease/trait risks in genome wide association studies.

Concurrent Session – Network approaches in biomedical research
Monday, August 8, 2022
2:30 - 4:40 PM
Logan

NetCellMatch: Multiscale Network-Based Matching of Cancer Cell Lines to Patients Using Graphical Wavelets

Neel Desai[1], Jeffrey Morris[1] and Veera Baladandayuthapani[2]

[1]Division of Biostatistics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104
[2]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109

Abstract
Cancer cell lines serve as model in vitro systems for investigating therapeutic interventions. Recent advances in high-throughput genomic profiling have enabled the systematic comparison between cell lines and patient tumor samples. The highly interconnected nature of biological data, however, presents a challenge when mapping patient tumors to cell lines. Standard clustering methods can be particularly susceptible to the high level of noise present in these datasets and only output clusters at one unknown scale of the data. In light of these challenges, we present NetCellMatch, a robust framework for network-based matching of cell lines to patient tumors. NetCellMatch first constructs a global network across all cell line-patient samples using their genomic similarity. Then, a multi-scale community detection algorithm integrates information across topologically meaningful (clustering) scales to obtain Network-Based Matching Scores (NBMS). NBMS are measures of cluster robustness which map patient tumors to cell lines. We use NBMS to determine representative "avatar" cell lines for subgroups of patients. We apply NetCellMatch to reverse-phase protein array data obtained from The Cancer Genome Atlas for patients and the MD Anderson Cell Lines Project for cell lines. Along with avatar cell line identification, we evaluate connectivity patterns for breast, lung, and colon cancer and explore the proteomic profiles of avatars and their corresponding top matching patients. Our results demonstrate our framework's ability to identify both patient-cell line matches and potential proteomic drivers of similarity. Our methods are general and can be easily adapted to other 'omic datasets.

PPIGCF: A Protein-Protein Interaction Based Gene Correlation Filter for Optimal Gene Selection

Soumen Kumar Pati[1,*], Manan Kumar Gupta[1], Ayan Banerjee[2], Saurav Mallik[3,*], Zhongming Zhao[3,4,*]

[1]Dept. of Bioinformatics, Maulana Abul Kalam Azad University of Technology, Haringhata-741249, WB, India; soumenkrpati@gmail.com, mownon89@gmail.com
[2]Dept. of Computer Sc. & Engineering, Jalpaiguri Govt. Engineering College, Jalpaiguri-735102, WB, India; ab2141@cse.jgec.ac.in

[3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; sauravmtech2@gmail.com
[4]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; zhongming.zhao@uth.tmc.edu

*Correspondence: zhongming.zhao@uth.tmc.edu, soumenkrpati@gmail.com, sauravmtech2@gmail.com 13

Abstract
Biological data at the omics level is highly complex, which requires powerful computational approaches for identifying significant intrinsic characteristics in order to further search informative markers involved in the phenotype of the study. In this article, we propose a novel dimension reduction technique, Protein-Protein Interaction based Gene Correlation Filtration (PPIGCF), which builds on Gene Ontology (GO) and protein-protein interaction (PPI) structures for analyzing microarray gene expression data. PPIGCF first extracts the gene symbols with their expression from the experimental dataset and then classify them based on GO Biological Process (BP) and Cellular Components (CC) annotations. Every classification group inherits all the information of their CC corresponding to the BP to establish the PPI network. Then, the gene correlation filter (regarding their gene rank and proposed correlation coefficient) is computed on every network and eradicates a few weakly correlated genes connected with their corresponding network. The PPIGCF finds the information content (IC) of the rest of the genes related to the PPI network and takes only the genes having the highest IC value. The satisfactory results of the PPIGCF will be used to prioritize genes as significant. We performed comparison with recent methods to demonstrate its efficiency.

Cancer Classification from Gene Expression Using Graph Attention Network

Sheikh Muhammad Saiful Islam[1], Ziqian Xie[1], Degui Zhi[1]

[1]School of Biomedical Informatics University of Texas Health Science Center at Houston, Houston, TX 77030, USA. Sheikh.Muhammad.Saiful.Islam@uth, Ziqian.Xie@uth, Degui.Zhi@uth}.tmc.edu

Abstract
The gene expression data from cancer patients has the potential to screen cancer very rapidly, saving critical time. Graph neural network, a kind of deep learning model, has the capability to model the gene expression while taking gene-gene interaction. We trained a graph attention network on gene expression profile (RNASeq) data from the cancer genome atlas program (TCGA) to differentiate 33 types of cancers. We have achieved 97.09% overall accuracy and 97.05% F1-score, outperforming existing models. In addition, we identified the cancer-type specific gene-gene interaction using calculated edge importance from GNNExplainer. We found that except for highly occurring gene-gene interactions among few genes, most cancers have unique significant gene-gene interactions. Our approach can be easily adapted to modeling other kinds of Omics datasets.

Identify Potential Driver Genes for PAX-FOXO1 Fusion-Negative Rhabdomyosarcoma Through Frequent Gene Co-expression Network Mining

Xiaohui Zhan[1], Yusong Liu[2], Asha Jacob Jannu[5], Shaoyang Huang[3], Bo Ye[1], Wei Wei[1], Pankita H.Pandya[4], Xiufen Ye[2], Karen E. Pollok[4], Jamie L. Renbarger[4], Kun Huang[5], Jie Zhang[6*]

[1]Department of Bioinformatics, School of Basic Medicine, Chongqing Medical University, China
[2]College of Intelligent Systems Science and Engineering, Harbin Engineering University.
[3]Carmel High School, Carmel, Indiana.
[4]Department of Pediatrics, Indiana University, School of Medicine
[5]Department of Biostatistics and Health Data Science, Indiana University, School of Medicine
[6]Department of Medical and Molecular Genetics, Indiana University, School of Medicine

*Corresponding author

Abstract
Rhabdomyosarcoma (RMS) is a soft tissue sarcoma usually originated from skeletal muscle. Currently, RMS classification based on PAX–FOXO1 fusion is widely adopted. However, compared to relatively clear understanding of the tumorigenesis in the fusion positive RMS, little is known for that in fusion-negative RMS (FN-RMS). Here we explored the molecular mechanisms and the driver genes of FN-RMS through frequent gene co-expression network mining (fGCN) and differential copy number (CN) and expression analysis on multiple RMS transcriptomic datasets. Among the 50 fGCN modules, five are differentially expressed between fusion status. A closer look showed 20% of Module 2 genes are concentrated on chromosome 8 cytobands. Upstream regulators such as MYC, YAP1, TWIST1 were identified for the fGCN modules. Further analysis in a separate dataset confirmed that, comparing to FP-RMS, 59 Module 2 genes show consistent CN amplification and mRNA overexpression, among which 28 are on the identified chr8 cytobands. Such CN amplification and nearby MYC and other upstream regulators (YAP1, TWIST1) may work together to drive FN-RMS tumorigenesis and progression. Up to 43.1% downstream targets of Yap1 and 45.8%
of the targets of Myc are differentially expressed in FN-RMS vs. normal comparisons, which also confirmed the driving force of these regulators.

Keywords: Fusion negative RMS (FN-RMS), Fusion positive RMS (FP-RMS), frequent co-expression network (fGCN), copy number alteration, upstream regulator

---

A Boolean network analysis of Parkinson's Disease genes and mitochondrial metabolic factors influencing Oxidative phosphorylation

Baby Kumari[1#], MD Zainul Ali[1#], Pankaj Singh Dholaniya[1*]

[1]Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad, Telangana - 500 046, India
#Equal authorship

*Corresponding Author:
Pankaj Singh Dholaniya, Ph.D.
Assistant Professor
S-46, Department of Biotechnology and Bioinformatics
School of Life Sciences, University of Hyderabad, Hyderabad 500046 India
Ph: (O) 91-40-23134591
Cell: 91-7799582919
Email: pankaz@uohyd.ac.in

ORCiD:
MD Zainul Ali 0000-0001-8001-8339
Baby Kumari 0000-0003-1340-7419
Pankaj Singh Dholaniya 0000-0001-8615-9861

Abstract
Background: Loss of dopaminergic neurons (DNs) is the primary cause of Parkinson's disease (PD). DNs have high energy demand, and oxidative phosphorylation pathway (OxPhos) is responsible for most of the ATP production, which is intricately linked with a few familial PD genes. Alterations in these genes promote SNCA aggregation and affect mitochondrial metabolic factors (MMFs). These MMFs are considered as key players in determining the functioning of OxPhos.
Objective: The detailed mechanism of regulation of the MMFs by PD-related genes involved in OxPhos is yet to be unveiled. In this study, we constructed a Boolean network that explains the role of these PD genes on OxPhos, MMFs, and the participation of intermediary components.
Methods: We performed GO analysis and literature survey that gives a list of familial PD genes which are implicated in OxPhos. The mechanisms of action of these genes and their interactions have been studied using Boolean network analysis.
Results: The Boolean model demystifies these PD genes' normal and pathological function and their effects on MMFs. It also explains probable mechanistic detail of the compensation pathway of a PD gene upon dysfunction of another PD gene in OxPhos. The model also suggests essential PD genes (DJ1, PARKIN, MNRR1, and LRRK2) and Ca2+ ions concentration as most crucial MMFs, whose pathological state will perturb the network substantially.
Conclusion: The function of these genes is interlinked, and changes in the PD genes affect MMFs which dysfunctions OxPhos that leads to Parkinson's disease.

Keywords: Parkinson's Disease, Oxidative Phosphorylation, Boolean Network, Derrida plot, Mitochondria, Neurodegeneration.

Concurrent Session – Artificial Intelligence on Big Data: Promise for Early-stage Trainees
Monday, August 8, 2022
2:30 - 4:40 PM
Rittenhouse

White blood cells and obesity: A Mendelian randomization study

James Yang[1], Yitang Sun1, Kaixiong Ye[1,2]

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

Abstract
Obesity is a global public health problem and its prevalence is increasing. The counts of white blood cells (WBC) have been previously associated with obesity, but the causality remains elusive. Establishing the causal roles of WBC in obesity not only enhances our understanding of the etiology of obesity but also provides preventative and therapeutic targets. This study aimed to investigate the causal association between WBC and obesity using Mendelian randomization (MR) analysis. Genetic instruments of 20 WBC traits were selected from three genome-wide association studies (GWAS, sample sizes ranging from 169,219 to 562,243 European individuals). Genetic associations of these genetic instruments with obesity were extracted from previous GWAS of obesity with large samples (n = 50,364 to 98,697). Our primary MR analysis used the inverse variance weighted (IVW) method under a multiplicative random-effects model. We evaluated the presence of heterogeneity with the Cochran Q statistic and the presence of horizontal pleiotropy with the MR-Egger intercept test. We further performed sensitivity analyses with the MR-Egger method, the weighted-median method, and the weighted-mode method. We found that one standard deviation increase in WBC is associated with lower risks of class I obesity (odds ratio (OR): 0.85; 95% confidence interval (CI): 0.78, 0.93; P = 5.04 × 10−4), class II obesity (OR: 0.71; 95% CI: 0.63, 0.81; P = 2.36 × 10−7), class III obesity (OR: 0.63; 95% CI: 0.51, 0.78; P = 3.44 × 10−5), and overweight (OR: 0.91; 95% CI: 0.85, 0.97; P = 3.34 × 10−3). These observations were robust to pleiotropic effects and supported by sensitivity analyses. In summary, our MR study supports the protective role of higher WBC in reducing the risk of obesity. Future studies are warranted to elucidate the mechanistic connections between WBC and obesity and to evaluate the usage of WBC in preventing and treating obesity.

Keywords: Genetic epidemiology, Mendelian randomization, White blood cells, Obesity

---

CoMutDB: The Landscape of Somatic Mutation Co-occurrence in Cancers

Chaoyi T. Zhang[1], Yan Guo[1]

[1]Department of Internal Medicine, Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM, 87109, USA.

Abstract
Motivation: One of the most important elements in carcinogenesis is somatic mutation. Somatic mutation co-occurrence has been proven to have a profound effect on tumorigenesis. While some studies have discovered links between co-mutations and clinical outcomes, a centralized resource dedicated to co-mutations in cancer is still lacking.
Results: Using multi-omics data from over 30,000 subjects and 1,747 cancer cell lines covering 78 distinct cancer types, we present the Cancer co-mutation database (CoMutDB), the most comprehensive resource devoted to describing cancer co-mutations and their characteristics.
The CoMutDB database and its web portal were developed using modern database and programming languages including MySQL, HTML, PHP, and JavaScript. The web portal of CoMutDB can accept users' queries with real-time computation. The CoMutDB can be queried by parameters including cancer type, tissue site, gene name, the number of co-mutated genes (up to 4), and mutation type (silent, nonsilent, and all). CoMutDB integrates survival data, drug sensitivity, and other clinical characteristics (age, sex, and stage) with the co-mutation status variable, and infers the statistical significance of the association between co-mutation and the various clinical characteristics.
Conclusion: CoMutDB is the first resource dedicated to describing co-mutation characteristics in cancers. Comprehensive co-mutation analyses revealed that certain pairs of co-mutations have greater prognosis prediction power than their solo components. Our case study further confirmed that detailed examinations of co-mutations can identify mechanisms that cooperate in tumorigenesis.
Availability: http://www.innovebioinfo.com/Database/coMutation/TCGA.php.

Keywords: cancer, somatic mutation, multi-omics, database, co-mutation, PHP

Identifying Putative Causal Links between MicroRNAs and Severe COVID-19 Using Mendelian Randomization

Chang Li[1], Aurora Wu[2], Kevin Song[3], Jeslyn Gao[4], Eric Huang[5], Yongsheng Bai[6,7], and Xiaoming Liu[1]

[1]USF Genomics & College of Public Health, University of South Florida, Tampa, FL 33612, USA
[2]Emma Willard School, Troy, NY 12180, USA
[3]Credit Suisse, New York, NY 10010, USA
[4]Simsbury High School, Simsbury, CT 06070, USA
[5]James E. Taylor High School, Katy, TX 77450, USA
[6]Next-Gen Intelligent Science Training, Ann Arbor, MI 48105, USA
[7]Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197, USA

Abstract
The SARS-CoV-2 (COVID-19) pandemic has caused millions of deaths worldwide. Early risk assessment of COVID-19 cases can help direct early treatment measures that have been shown to improve the prognosis of severe cases. Currently, circulating miRNAs have not been evaluated as

canonical COVID-19 biomarkers, and identifying biomarkers that have a causal relationship with COVID-19 is imperative. To bridge these gaps, we aim to examine the causal effects of miRNAs on COVID-19 severity in this study using two-sample Mendelian randomization approaches. Multiple studies with available GWAS summary statistics data were retrieved. Using circulating miRNA expression data as exposure, and severe COVID-19 cases as outcomes, we identified ten unique miRNAs that showed causality across three phenotype groups of COVID-19. Using expression data from an independent study, we validated and identified two high-confidence miRNAs, namely, hsa-miR-30a-3p and hsa-miR-139-5p, which have putative causal effects on developing cases of severe COVID-19. Using existing literature and publicly available databases, the potential causative roles of these miRNAs were investigated. This study provides a novel way of utilizing miRNA eQTL data to help us identify potential miRNA biomarkers to make better and early diagnoses and risk assessments of severe COVID-19 cases.

Keywords: microRNA, SARS-CoV-2, COVID-19, biomarker, Mendelian randomization

Citation
Li, C., Wu, A., Song, K., Gao, J., Huang, E., Bai, Y., & Liu, X. (2021). Identifying Putative Causal Links between MicroRNAs and Severe COVID-19 Using Mendelian Randomization. *Cells*, *10*(12), 3504. https://doi.org/10.3390/cells10123504

A Deep Learning Model for Ancestry Estimation with Craniometric Measurements

Kevin Ma[1], Xiaoming Liu[2]

[1,2]College of Public Health, University of South Florida, Tampa, FL, USA

Abstract
Ancestry estimation from human skeletal remains is the process of estimating a person's geographic origin by comparing their craniometric measurements against the craniometric measurements of people indigenous to the area. Ancestry estimation based on skeletal measurements is still a significant component in modern forensic anthropology studies. Although there are computational tools for ancestry estimation, the accuracy and performance of such tools are not very good in actuality. Two of the most popular ancestry estimation software, Fordisc 3.1 and AncesTrees, use canonical variates analysis and the random forest algorithm, respectively, to match craniometric data to a geographic location. However, a deep learning approach to this process was never utilized. In this paper, we evaluated the practicability and efficacy of a deep learning method in cranial ancestry estimation based on Howells craniometric data. Our paper analyzed the cranial data of 2,524 individuals from the Howells main datasets and 468 from the Howells test datasets. Individuals with 82 craniometric measurements in the Howells datasets were grouped into six categories based on geographic ancestry: African, East Asian, Native American, Austro-Melanesian, Polynesian-Micronesia, and European. Data from the Howells dataset was first transformed through missing value imputation, standardization, normalization, and one-hot encoding. After the data engineering process, a feedforward neural network (FNN) model was created using data from the Howells datasets with 30 craniometric measurements. The feedforward neural network model was trained on 80% of the data from the Howells main datasets,

validated on 10% of the data from the Howells main datasets, and tested on another 10% of the data from the Howells main datasets. The model had a prediction accuracy of 80.6% for all six categories for 10% of the Howells main dataset. When compared with the ancestry estimation programs AncesTrees and Fordisc 3.1 using the Howells test dataset, the performance of the feedforward neural network deep learning method was generally similar and showed better performance for some ancestral groups. The developed feedforward neural network model performed significantly better for East Asian and African groups and relatively worse for the Austro-Melanesian and Native American groups. When tested against the whole Howells main dataset, the developed feedforward neural network model had an accuracy of 62.4%, in comparison to 62.2% (Fordisc 3.1) and 63.5% (AncesTrees).

Y. Dong et al., "A Deep Learning Model for Ancestry Estimation with Craniometric Measurements," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 3350-3357, doi: 10.1109/BIBM52615.2021.9669742

Keywords: ancestry estimation, craniometric measurement, deep learning

---

Condition-specific Gene Co-expression Network Analysis Reveals Copy Number Variations Associated with KRAS Mutation Status in Colon Cancer

Shaoyang Huang[1,2], Chi Zhang[2*]

[1]Carmel High School, Carmel, IN 46032
[2]Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, US 46202
*To whom correspondence should be addressed. Chi Zhang (czhang87@iu.edu)

Abstract

Background: About 30-40% of colon cancer patients carry KRAS mutations. These patients usually have poor prognosis including more aggressive tumors and shorter survival time. In order to understand the impact of KRAS mutations in the tumor tissues, especially on the relationships among gene expression levels, we carried out condition-specific gene co-expression network analysis (GCNA). Our goal is to identify co-expressed gene modules in colon cancers that are specific to patients with KRAS mutations or without KRAS mutations. By inspecting the condition-specific gene modules, we aim at inferring new insights related to important biological processes and structural variations associated with KRAS mutations.

Methods: We obtained gene transcriptomic data (RNA-seq) for the TCGA COAD cohort from UCSC Xena data portal. Among the 512 patients after initial filtering, 164 have KRAS mutations (with amino acid changes) with potential biological impact and 348 have no KRAS mutations. GCNA were carried out for each group (KRAS mutant vs non-mutant) separately using the lmQCM algorithm with the same parameters implemented in Matlab. Gene modules with at least 20 genes were kept for each group. Gene modules in one group containing less than 20% overlap (measured using Jaccard index) with all gene modules in the other group is considered condition specific. Enrichment analysis for these condition specific genes were carried out using TOPPGene for further interpretation. For those gene

modules with genes that are significantly enriched on specific cytobands on chromosomes, uniquely enriched cytobands were also identified for each group.

Results: We identified 25 of co-expressed gene modules for the KRAS mutant group and 30 for the non-mutant group. Among them 11 are specific to the KRAS mutant group and 14 are specific to the non-mutant group. Enrichment analysis on these condition-specific gene modules identified 9 (out of 11) are significantly enriched on different cytobands for the KRAS mutant group while all 14 are significantly enriched on different cytobands for the non-mutant group. Among the enriched cytobands, 5q12.3, 17p13.2, and 4p16.3 are unique for the KRAS mutant group and 20q13.33, 8q24.3, Xq21.2, Xq28, and 18q21.1 is unique for the non-mutant group. Since the cytobands that are enriched with co-expressed genes often imply existence of larger variations in gene copy numbers, we selected gene from these cytobands and inspected their copy number readings between the two groups using cBioPortal. We observed consistently larger range of copy number variances in the group to which the original gene modules were specific. For instance, as shown in Figure 1, the important colon cancer gene APC is on chr5q21, which is a cytoband specifically enriched in the KRAS mutant group and its copy number readings have a significantly larger variation (F-test $p=0.00272$) in this patient group comparing to the non-mutant group while GATA5 from 20q13.33 shows an inversed pattern ($p=0.00027$).

Discussion and conclusion: GCNA is a powerful tool that not only can infer gene-gene relationships but also can contribute to the detection of copy number variances. We observed that there are more cytobands that are specifically enriched in the non-mutant group, suggesting that there are different mechanisms related to genome stability in this group. These cytobands are also associated with important cancer genes. For the KRAS mutant group, the only uniquely enriched cytoband contains the important colon cancer gene APC while for the non-mutant group, the enriched cytobands cover important genes such as MYC (8q24) and TP53 (17p21). Interestingly chromosome X is highly enriched in the non-mutant group, suggesting the gender as a factor. Overall, our analyses have led to new hypotheses regarding the relationships between KRAS mutation status in colon cancer and the other genetic variations and biological processes. Further analysis regarding the combinatorial effects of these mutations on patient outcomes will shed light on developing precision treatment for the patients.

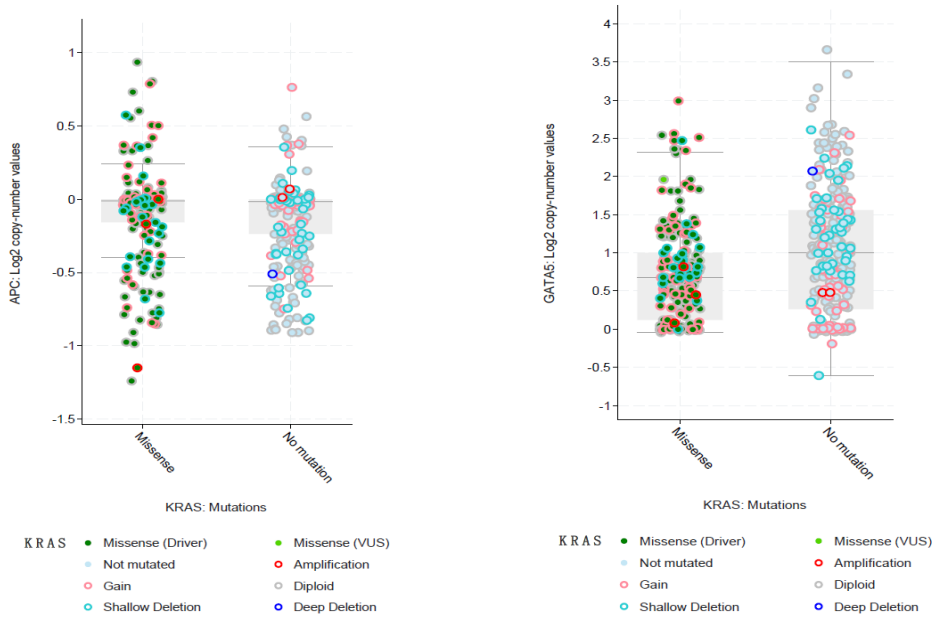


Figure 1. Examples of copy number readings of selected for colon cancer patients (from TCGA COAD cohort) with missense mutations in KRAS versus those without KRAS mutations. APC (left) is on 5q12.3 that is specifically enriched in the KRAS mutant group while CHD4 is on 20q13.33 that is specifically enriched in the non-mutant group.

Computational modeling of cell type specific metabolic rate of glucose flow and glutaminolysis in cancer microenvironment

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

[1]Center for Computational Biology and Bioinformatics, [4]Department of Biostatistics, [5]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN;
[2]Carmel High School, Carmel, IN;
[3]Park Tudor School, Indianapolis, IN. (+equal contribution)

Abstract

Glutaminolysis has been considered as a major carbon and energy source that promote the proliferation of cancer cells. Recent studies reported immune cells trend to use more glucose while cancer cells use more glutaminolysis for ATP production. The observations made on cell line or limited mouse orthotopic models may not reflect the general metabolic shifts in real human cancer tissue. In this study, we conducted a computational analysis to characterize the flux and distribution of glucose and glutamine metabolism in cancer, including production of lactate (glycolysis), TCA cycle, nucleic acids synthesis, glutaminolysis, glutathione and amino acids synthesis, in multiple cancer types and normal tissue, by using TCGA tissue transcriptomics data and 10 scRNA-seq data of human cancer and mouse orthotopic tumors. Our analysis confirms the increased influx in glucose uptake and upper part of glycolysis and decreased upper part of TCA cycle in cancer cells. However, increased lactate production and second half of TCA cycle were only seen in certain cancer types. More interestingly, we did not see cancer tissues have highly shifted glutaminolysis comparing to normal tissue samples

in human cancer. Although cancer cells have highest glutaminolysis rate compared to stromal and immune cells in tumor microenvironment (TME), we did not observe stromal and immune cells have higher glucose consumption than cancer cells. Further analysis illustrates immune cells may have higher potential in glucose uptake (by glucose transporters), while their downstream consumption of glucose is much lower than cancer cells. A systems biology model of metabolic shifts and competition is further developed.

We have recently developed a novel computational method, namely single-cell Flux Estimation Analysis (scFEA) to estimate sample-wise metabolic flux by using tissue or single cell transcriptomics data. We reconstructed a metabolic map that include glycolysis and key branches of glycolysis, TCA cycle, glutaminoloysis, import of glucose, glutamine and glutamate, in-/out-flux of glutamine and glutamate involved in other metabolic reactions, in human and mouse. We reconstructed the curated pathway into a factor graph, in which each variable represents a metabolic module, and each factor represents one intermediate substrate. The reconstructed factor graph includes 27 metabolic modules, 165 genes in mouse and 176 genes in human.

Our key results include, 1. In both human and mouse, by using flux analysis, in all analyzed data, cancer cells have more glucose and glutamine metabolism rate than myeloid cells and T cells; 2. Cancer cells have higher or even glucose uptake rate than myeloid cells and T cells; 3. Both glutamine and glucose metabolism highly associated with cell proliferation, the former one fuel amino acids biosynthesis while glucose contribute to DNA biosynthesis; 4. Myeloid cells has higher proliferation rate in human TME than mouse TME; 5. In mouse, as myeloid cells does not proliferate, the glutamine uptake is lower than cancer cells. While in human cells, myeloid cells may take substantial amount of glutamine, majorly to fuel cell proliferation; 6. In human and mouse, all cell types have relative evenly expressed glucose transporter; 7. In human TCGA data, glucose transporter positively correlated with immune cell proportions while glucose metabolism positively correlated with cancer cell level; 8. In human TCGA data, different cancer types have highly varied glutaminolysis rate and lactate production rate, but their TCA cycle rate are relatively even; 9. In both human and mouse, majority of glutamine fuel biosynthesis of other amino acids than glutaminolysis.

Computational modeling of cell type specific metabolic rate of Branched Chain Amino Acids

Kevin Hu[1,2+], Alex Lu[1,3+], Shaoyang Huang[1,2+], Grace Yang[1,2+], Noah Meroueh[1,2+], Haiqi Zhu[1,3], Pengtao Dang[1,3], Sha Cao[1,4*], Chi Zhang[1,5*]

[1]Center for Computational Biology and Bioinformatics, [4]Department of Biostatistics, [5]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN;
[2]Carmel High School, Carmel, IN;
[3]Park Tudor School, Indianapolis, IN. (+equal contribution)

Abstract

Both cancer and Alzheimer's disease have altered branched chain amino acids (BCAA) metabolisms. Recentstudies identified the ratio of leucine, isoleucine, and valine metabolism determines the disease characteristics. In this study, we reconstructed the BCAA metabolic modules to enable the first computational capability to estimate fluxome of BCAA metabolism. Our analysis characterized the flux distribution of BCAA metabolism including synthesis of branched chain fatty acids (BCFA) and production of acetyl-CoA, succinyl-CoA and propanoyl-CoA. We identified cancer trends to have a

relatively lower level of isoleucine metabolism to succinyl-CoA and propanoyl-CoA and save most isoleucine for BCFA biosynthesis. On the other hand, most leucine and valine were converted to succinyl-CoA and acetyl-CoA. In AD, most of the BCAA metabolisms happen in astrocyte, neuron and oligodendrocyte progenitor cells with different characteristics. Distinct cell type specific metabolic status of BCAA has been observed.
This study was conducted by five high school students from the STEM program. They established a context mining based approach to identify and annotate high quality single cell RNA-seq data of human cancer microenvironment. 60 data sets were identified and processed for the study of this study and other projects to study metabolic changes in cancer.

Interactive Data Collection Using CARLA and OpenCDA for Reinforcement Learning

Alan Chen[1,2], Joseph Clemmons[1], Umar Jamil[1], Ashley Land[3], Sara Ahmed[1], Yu-Fang Jin[1]

[1]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX, USA;
[2]Westlake High School, Austin, TX, USA;
[3]Department of Software Engineering, St. Mary's University, San Antonio, TX, USA

Abstract
Autonomous vehicles (AVs) have attracted significant research efforts driven by an over $9 billion market value by 2027. Autonomous driving has been considered to be the future solution to reducing crash rates and traffic flow while ensuring drivers' safety and enabling many possibilities. This project sought to create a virtual driving scenario, by simulating a live road situation, to collect interactive training trajectories for inverse reinforcement learning (RL).
To establish this virtual environment, CARLA simulator (http://carla.org/), SUMO (https://www.eclipse.org/sumo/), and OpenCDA (https://opencdadocumentation.readthedocs.io/en/latest/md_files/introduction.html#) were deployed. OpenCDA was used to interface with CARLA and SUMO and enabled real-world scenarios, in-simulation communication between vehicles, and a logic flow. SUMO's python API was utilized to develop XML files for different traffic environments, such as the number of vehicles in the system, their initial locations, initial speed, and final locations during traffic generation. A scenario was established, comprising of a platoon with connected AVs (CAV) driving on the main road and an ego AV(merger) merging onto the main road. A python script was developed that randomized the yaml file's parameters such that the characteristics of the CAVs in a platoon and the speed of the platoon and ego AV varied during each episode. At the end of each episode, data about the platoon's characteristics were extracted. This open framework is capable of including additional behaviors of the ego AV for more complicated training and provides flexibility during testing and the recording of the platoon and ego AV's characteristics.
Using a map of a two-lane highway, the parameters for each CAV and the ego AV were randomized for over 50 episodes. Three types of interactions were listed as follows. Merging was defined as the ego AV joining the platoon between two AVs. Joining was defined as the ego AV joining the platoon as a follower at the end. Failure was defined as the ego AV missing the

platoon. The table, stored as a CSV file, included the randomized speed of the ego AV and platoon, merging time, and interaction types. Correspondingly, the characteristics of each simulation were recorded. A total of 1,000 simulations were performed to collect data as input into RL training.
The project provided a convenient and flexible framework for collecting training data on a specific scenario. The data values retrieved from the CSV file are vital to achieving an increased success rate with inverse reinforced learning.

Keywords: Data Collection, Reinforcement Learning, Autonomous Vehicle, CARLA, OpenCDA, SUMO

Optimal Charging Strategy for Electric Vehicles

R.J. Alva[1], Albert Zhang[2], Eugenia Cadete[1], Sara Ahmed[1], Yu-Fang Jin[1]

[1]UTSA Department of Engineering, San Antonio, TX, USA; 2Wissahickon High School, Ambler, PA, USA.

Abstract

Significance: With the rapid increase in the number of Electric vehicles (EVs) on the road, the charging schedule of EVs has a significant impact on the efficiency of the charging equipment and the reliability of the power grid. EVs charging at workplaces during the daytime causes extra loads to the power grid and the charging load profiles are embedded with spatial and temporal uncertainties, resulting in extra difficulty in scheduling the EV charging. In this study, optimal charging strategies were developed to schedule EV charging with objectives concerning charging stations efficiently with higher charging rates and mitigating the strain on the power grid by smoothing the duck curve.

Methods: Over 32,000 charging sessions from more than 121 charging stations located at three different workplaces with different public accessibilities were collected and analyzed to extract information including connecting/disconnecting time of a charging session, charging duration, instant charging current, and total energy requested for a charging session. Charging current and energy were normalized with respect to the maximum value of the variable in the dataset. The highest probability of charging occupancy was determined and used to schedule EV charging. Scheduling of the EV charging was achieved by determining the charging rate in a session with four different objectives, considering charging efficiency with linear combination and nonlinear weight of charging rates, smoothing the duck curve, and a combination of different objectives. A duck curve was generated using temporal profiles of power grid loads. Quadratic programming was applied to optimize the objective functions under multiple constraints concerning the availability of charging stations, the relationship between delivered and requested energy, and peak loads of the power grid. A graphical user interface was also developed to illustrate the real-time power consumption, charging schedules, and availability of CS.

Results and Conclusion: Four different charging scheduling strategies were developed and simulated. With reduced power (75%) delivered to EVs, all 4 strategies reduced the average duty percentile significantly with an average duty percentile of 52.3% by smoothing the duct curve and 22.4% by deploying nonlinear weighted charging efficiency. In addition, scheduled charging sessions considering the smoothness of the duck curve delivered the power to EVs with fewer (50% less)

charging peaks than the scheduled session considering only charging efficiency. Fewer peaks in the charging session potentially benefit the battery of EVs.

Keywords: Optimal Charging, Electric Vehicles, Duck Curve

Artificial Intelligence-Based Navigation System for Automated Wheelchairs

Jeffrey Wang[1], Kevin Liu[2], Yu-Fang Jin[3]

[1]Keystone School, San Antonio, TX, USA.
[2]University of Western Ontario, London, ON, Canada.
[3]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX, USA

Abstract
It's reported that 3.6 million people in the United States above the age of 15 use wheelchairs, most of which are powered wheelchairs (PWC), to assist in moving from place to place. However, one issue with PWCs is that those with severe cognitive, motor, or sensory issues cannot operate them fully. Further, PWC users report difficulty in navigation through indoor public spaces. Vision-based navigation is in demand for automatic indoor navigation of PWC. With the superior performance of artificial intelligence (AI) in image processing, the goal of this project is to provide an accessible and affordable edge-based AI solution that can be used on users' existing PWCs.
As most PWCs come with user controls such as a joystick and remote control without any navigation aids, with a mobile app that communicates with a joystick or remote control, the ResNet50 model was adopted to process images captured by the built-in cameras on mobile phones. This edge-based AI navigation system is developed with a modified ResNet50 model. A total of 2,000 images captured by mobile devices installed on PWC were collected and labeled based on the motion direction of a PWC. Processed images with 240 by 240 resolution served as inputs to the ResNet50 with 4 steering controls as outputs: forward, left, right, and stop. This module seeks to find a collision-free path for PWC. About 80% of images in different locations were used to train the model and the rest images were used to test the model. A 95% accuracy was achieved in the test data sets. This AI-based navigation system could effectively turn a PWC that already has remote control capabilities into a "smart" PWC without forfeiting its original capabilities.

Detecting Unattended Baby in Car-seat as a New Vehicle Safety Feature

Jerry Guo[1], William Xiao[1], Yu-Fang Jin[2]

[1]Louis D. Brandeis High School, San Antonio, TX USA
[2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX USA

Abstract

Significance: Each summer, about 38 unattended children die inside a stalled car due to heatstroke in the United States, leading to a significant loss for these families. However, no effort has been dedicated to addressing this issue to the best knowledge of the authors. Successful applications of artificial intelligence in image processing inspired our effort to develop an onboard application to detect unattended babies in car seats and link the detection results with car keys or locks.
Methods: We collected 600 annotated vehicle interior images with (positive samples) and without (negative samples) a baby in the car seat, 300 images for each category. The negative samples also include images of stuffed animals in a car seat with varied fuzziness. All pictures were augmented by randomly cropping at a scale of 128 by 128 pixels. We applied the "fastai" Python package with a pretrained ResNet34 framework to detect whether a "baby" sat in a car seat or not. The model was trained on a Linux server equipped with RTX 3080 12G GPU. A 5-fold cross-validation was conducted for the performance validation.
Results: Our model reaches 0.97 for all performance metrics of accuracy, precision, recall, and F1-score, after merely 50 epochs of training cycles. This balanced cross-all performance metric can be attributed to the balanced training dataset. In addition, the model development code and data set for this project is sharable at (https://github.com/s698667/ICIBM_AI_baby_in_carseat).
Conclusion: Given the limited amount of training data, we noticed that detection accuracy is better with the front-view images of a baby in a car seat than the results obtained from side-view images. A more comprehensive investigation has been planned to include more side-view images and fuzzy images to further improve the performance of the model. We expect this application will work with motion detector and bring about a tremendous social-economic impact on vehicle safety improvement.
Acknowledge: NSF EEC-2051113, USDOT Transportation Consortium of the South Central States (TRAN-SET)-21034 and -21049.

Keywords: Fast.ai, image classification, transfer learning, ResNet34, Unattended Baby Detection

---

Decentralized Collision-Free Trajectory Planning for Autonomous Vehicles Using Reinforcement Learning

Joseph Clemmons[1], Umar Jamil[1], Alan Chen[1,2], Ashley Land[3], Sara Ahmed[1], Yu-Fang Jin[1]

[1]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX, USA;
[2]Westlake High School, Austin, TX, USA;
[3]Department of Software Engineering, St. Mary's University, San Antonio, TX, USA

Abstract
Using reinforcement learning (RL) to guide an autonomous vehicle (AV) has been a booming research area, due to the advantage of investigating interactions among multiple agents and the environment. The goal of this project is to develop an RL algorithm to guide an AV with a collision-free trajectory and verify the results in CARLA simulator.
To reach the goal, a framework including CARLA, CARLA's Python API, and Tensorflow was established. A CNN-based q-learning algorithm was developed in Python. Images from an RGB camera were input into a CNN model. Outputs of the CNN model represented the q-values of the three

steering actions (Left, Straight, and Right) in the q-learning algorithm. The policy of choosing the highest q-value action in the current state or choosing a random action was determined by a policy score: the higher the score was, the more likely a random action occurred. An action was given a positive reward for each step without a collision, determined by a collision sensor in CARLA, while maintaining the desired speed. The CNN-based q-learning model was trained with 1,000 images and guided an AV with a collision-free path.

Due to the delay in image processing, a search-based q-learning model with a look-up table was developed. States of this model were defined as: 1) the distance, δ, between the AV and its desired position, 2) the angle, α, between the AV's orientation and its motion direction (a vector from its current position to the desired position), and 3) the angle, θ, between an AV's orientation and the heading of the road. Five steering actions were defined as straight, fully powered left and right, and half-powered left and right. The same policy was adopted as the CNN-based q-learning. The reward was defined as, $¼ [2(\cos\alpha-0.5) + 2(\cos\theta-0.5) + (C-\delta)/C + \tau] + \rho$, where C is a predefined distance range, τ is the time penalty forcing the AV to reach its destination and ρ is a reward for reaching its waypoint. To train the q-learning model, the lookup table was randomly initialized and updated, based on the current state and actions taken at each iteration. This q-learning model guided an AV from an initial position to the desired destination while staying in a particular lane. The search-based q-learning algorithm was able to reach the waypoint with a 97% success rate. Integrating the CNN-based and search-based qlearning models is planned for future research.

Keywords: Reinforcement Learning, Autonomous Vehicle, CARLA

Developing a Sustainable Low-maintenance Traffic Crash Risk Notification System

Seth Klupka[1], Tulan R. Sampath Bandara[2], Paul Morton[2], Mimi Xie[1], Yu-Fang Jin[2]

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

Abstract

Significance: Each year, 90% of the roughly 36,000 traffic-related deaths in the U.S. are the result of human errors according to the Automobile Association of America. Keeping drivers alert with early safety notifications on potential risks is important to reduce human errors and improve public safety. The goal of this project is to develop a sustainable low-maintenance realtime traffic crash risk notification system.

Methods: The notification system consists of a Raspberry Pi 4 for risk prediction, two wireless communication modules ESP32 LoRa OLED, Arduino Mega 2560, LED matrix, and solar panels. A traffic risk level produced from the Raspberry pi 4 was transmitted via radio frequency (RF) through a LoRa antenna due to its low cost, range, and high energy transfer efficiency. The message transmitted was caught by a receiver (ESP32 LoRa OLED development board) and sent to an Arduino Mega 2560, then displayed on an LED matrix panel. Both subsystems are powered via a sustainable, low-maintenance source referred to as an energy harvester. The message was updated and displayed every 2 seconds which should be sufficient for real-time traffic crash risk notification. The communication frequency is also adjustable for further energy-saving.

Results: The whole signal transmission flow from the Raspberry Pi 4 to the LED matrix has been assembled and tested. The working range of the remote notification has been tested and calculated based on the GIS information. The estimated longevity of the system without sunlight is about 10 hours. **Conclusion:** The integration of the Raspberry Pi, wireless communication modules, and the LED matrix validated the concept of real-time notification of traffic crash risks predicted by the Raspberry Pi. The framework developed in this study will allow an edge-based prediction system using a Raspberry Pi.

Keywords: Traffic crash risk prediction; Deep learning; Energy harvesting.

Transfer Learning for Cancer Survival Prediction using Gene Expression Data

Ariel Lee[1], Jacob Buckle[1], Ricardo Ramirez[1], Yu-Fang Jin[2]

[1]Department of Engineering, Houston Baptist University, Houston, TX, USA.
[2]Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, USA.

Abstract
Over the past few years cancer survival rates have increased due to more advanced medicine and technology such as Next Generation Sequencing (NGS). NGS has made gathering RNA-Seq data much easier, leading researchers to develop personalized treatments for specific cancer types. Many attempts have been made to use Machine Learning and Deep Learning to extract features and information from the high dimensional RNA-Seq dataset to learning more on significant genes that can lead to early detection of cancer. In this paper Deep Learning was used on 13 of The Cancer Genome Atlas (TCGA) cancer datasets to predict a patient's risk score. From all the cancer types in TCGA only the cancers with significant information were used. The criteria being that the cancer dataset must contain more than 100 patients and more than 50 of those patients being non-censored. Using Cox-Survival analysis, four Deep Learning models were successfully developed, such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN) and Transfer Learning on the ANN and CNN models. The Transfer Learning model was developed by training the model on all other 12 cancer types except for the cancer type left out for fine tuning the model. This was done for all 13 cancer types. We compared all Deep Learning models to a baseline model, Cox-PH model, to verify performance. The Deep Learning models showed significant improvement in predicting the risks scores of patients over the original PH-model due to the complexity of cancer. However, Transfer Learning did not improve survival scores for all cancer types due to some cancer types being too unique compared to others.

Keywords: Transfer Learning, TCGA, Cancer, Deep Learning

Analysis for Traffic Crash Severity in Texas Using CRIS Database

Shuyu He[1], Taylor Jones[2], Tulan Sampath Bandara[2], Seth Klupka[2], Sara Ahmed[2], Samar Dessoukyand[3] and Mimi Xie[2]
[1]Columbia University, NY, USA;

[2]The University of Texas at San Antonio, San Antonio, TX, USA;
[3]Brandies High School, TX, USA;

Abstract
Road traffic accidents are the leading cause of death for people aged 5-29 in the United States (U.S.). Fatal crashes occur on collector roads (41% and 9% in rural and urban areas, respectively) and local roads (19% and 13% in rural and urban areas, respectively), and non-intersection roads (85% and 68% in rural and urban areas, respectively). Public databases containing crash information have been established, however, there is still a lack of thorough analysis of these data to gain a better understanding on what are the contributing factors to the severity of traffic crashes. Therefore, the goal of this project is to analyze the crash information stored in the Department of Transportation's (TxDOT's) Crash Records Information System (CRIS) database and extract the features of traffic crashes in Texas.

The non-sensitive crash information from 98,792 crashes in CRIS 2021 database contains multiple sheets as CSV files grouped into two-month periods. The 700MB data was not organized in an efficient way and the datasheet contains text-based annotation and are not suitable for data-driven analysis. An automatic information process package was programmed in Python to extract variables including crash ID, time and location of crashes, road surface conditions, weather conditions, environment light conditions, speed limit, the damage to property, parameters of the crash, number of people injured, the severity of injures, and the number of deaths involved in a crash. We programmed in R to create and visualize the correlation matrix based on the primary person's injury severity levels. We categorized the involved person's role in the accident, either as drivers, passengers, or pedestrians then performed correlation and regression analyses for each personnel category and 12 potential risk factors.

A total of 18 variables were identified as potential risk factors for crash severity based on a literature review. Annotations of the variables were quantified as -1 for unknown information, 0 for least severe/not affected, and 1 – 5 denoted an increased severity. The quantification of text annotation allows for further data-driven analysis. Our preliminary analysis showed that person injury severity, person airbag ejection, and person ejection were strongly correlated with each other. Interestingly, the number of persons involved in a crash was not related to the severity.

Keywords: Computational modeling of metabolic flux, Branched Chain Amino Acids, Cancer, Alzheimer's Disease

---

Cell-Type Identification with Single Cell RNA-Sequencing Data for Temporomandibular Joint Disorders

Mostafa Malmir[1], Savannah Lopez[2], Marlene Luo[3], Karen Lindquist[4], Sergei Belugin[4], Arman Akopian[4], Yidong Chen[5,6], Jinyan Li[7] and Yu-Fang Jin[2]

[1]Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, USA;

[2]Department of Computer Science, St. Mary's University, San Antonio, TX, USA;
[3]Westwood High School, Austin, TX, USA;
[4]Department of Endodontics, School of Dentistry, University of Texas Health San Antonio, San Antonio, TX, USA;
[5]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX, USA;
[6]Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX, USA;
[7]Department of Statistics and Data Management, University of Texas at San Antonio.

Abstract
Temporomandibular joint disorder (TMJD), examined in 40-60% of the population, is a serious health problem due to its recurrent nature and difficulty in managing the chronic pain of TMJD. Identifying types of neurons in TMJD is very important to better understand the disease and maximize therapeutic intervention. Single Cell RNA-sequencing (scRNAseq) has been widely adopted for biological research due to its capability to characterize expression profiles at single-cell levels. However, few efforts have been dedicated to investigating neurons at singlecell levels. In this study, we seek to identify cell types relating to the Trigeminal Ganglion (TG) with scRNAseq data. The cDNA libraries were constructed with 192 cells from 8 C57 mice and then sequenced, leading to a total of 15,677 features in the dataset. Both Cell-ID and Seurat were adopted to analyze this dataset. As a cluster-free approach, Cell-ID performed multiple correspondence analyses and projected both cells and genes in a common orthogonal space, allowing us to find genes close to each cell and extract gene signatures per cell based on gene-tocell distances. We combined two cell marker databases, CellMatch and Panglao, and found 88 different cell types, among which 62 cell types have more than 5 gene markers. The gene markers were used for cell type identification. We evaluated the similarity between closest genes to a cell and the gene markers of 62 cell types and assigned cell type to each cell based on the Pvalue of the similarity. As a comparison, the Seurat pipeline was also used to process the data and assigned cell type using the same gene markers. A total of 19 cell types were assigned for 157 cells and the rest 35 cells were not assigned to any cell type. Neuron cells identified include 39 trigeminal neurons, 30 neurons, 28 type I spiral ganglion neurons, 21 pan-gabaergic, 11 type II spiral ganglion neurons, 2 Type IC spiral ganglion neurons, and 2 interneurons. The type I spiral ganglion neurons had the highest confidence for cell type identification. Results obtained with Seurat showed fewer cell types identified and were sensitive to the parameters fordimensionality reduction and threshold for clustering. Cell-ID demonstrated the advantage of assigning cell types to each cell with a quantified confidence level and extracting gene signatures for each cell. Gene signatures of unassigned cells have the potential to discover new cell types or subtypes.

Keywords: Single-cell RNA seq, Cell Identification, TJMD

Retrieving Knowledge of Molecular Mechanisms from Literature Titles via an Event Extraction Approach

David Spellman[1], Jason Xiaotian Dou[1], Aaron Fangzheng Wu[1] and Yufei Huang[1]
[1]University of Pittsburgh, PA, USA

Abstract
Automatically extracting knowledge from a large number of literatures is an impactful and challenging problem in biomedical research. Most current approaches use the pipeline including named entity recognition and relation extraction. However, relation extraction mostly extracts primitive associations in the form of positive or negative associations, far from capturing the key knowledge in the literature that describes the regulatory mechanisms among biomedical entities, the knowledge fundamental to the understanding of biological processes. The diverse sentence structures and varying patterns of paragraphs that describe regulatory mechanisms significantly challenge the learning of this knowledge from abstracts or full articles. In contrast, titles are the most succinct summary of an article's key findings, mostly relating to regulatory mechanisms by their importance. While current knowledge extraction approaches usually focus on abstracts, few consider learning regulatory mechanisms from titles. In this work, we examine the problem of extracting the knowledge about molecular regulatory mechanisms from titles. To this end, we formulate the problem as an event extraction and adopt the GYGIE++ framework for its solution. To train this model, we exploit the literature on m6A mRNA methylation and constructed the first-ever event extraction training dataset focusing on molecular mechanisms. To construct this data, we meticulously defined the biological event with the trigger words such as "induce" and "inhibit" and designed a template that include 5 arguments that summarizes the knowledge as " Arg1 regulates/suppresses/promotes Arg2 of/in Arg3 by controlling/regulating Arg4 of/in Arg5". Preliminary experiments show promising results of our approach. We achieved mean argument identification precision as 0.82, mean argument identification recall as 0.72, mean argument identification F1 as 0.77, and mean argument classification F1 as 0.71 in the test data. In future work, we plan to expand our definitions of biomedical mechanism events and templates and conduct comprehensive evaluations. Ultimately, we plan to apply this method to the entire PubMed articles to construct the molecular mechanism knowledge graphs which can genuinely support biomedical knowledge discovery and downstream tasks like question answering and information retrieval. We acknowledge the support of the NSF REU program EEC-2051113.

Keywords: Molecular Mechanisms, Literature Titles, Event Extraction, Knowledge Discovery

---

Abstract ID: 89

Mapping Imaging Genetics Associations at Multiple Scales: A Study of Cortical Thickness Phenotypes for Alzheimer's Disease

Kevin Shen[1], Manu Shivakumar[2], Dokyoon Kim[2]

[1]Harriton High School, Bryn Mawr, PA, USA
[2]University of Pennsylvania, Philadelphia, PA, USA

Abstract
Introduction: Alzheimer's disease (AD) is a national research priority. With 5.8 million Americans affected, it has no available cure, and is a huge economic burden. To progress in drug development, effective strategies to understand the AD genetic mechanisms are needed. An emerging direction offering enormous opportunities to unravel biological pathways between genome, brain, and disease is using neuroimages as phenotypic traits to study AD genetics. Utilizing rich imaging genetic findings

from UK Biobank (UKBB), we performed a multiscale analysis of imaging genetic associations for AD causal genes and cortical thickness phenotypes.
Methods: Our analysis focused on 1) single nucleotide polymorphisms (SNPs) in 20 AD causal genes (https://adsp.niagads.org/index.php/gvc-top-hits-list/); 2) 68 regional cortical thickness quantitative traits (QTs) measuring neurodegeneration. Using UKBB SNP-QT summary statistics, we extracted all SNP-QT associations satisfying a user-specified significance level (e.g., $p \leq 0.05$) to form a fine-level SNP-QT association map. Next, we created an intermediate-level gene-QT association map. For each gene-QT pair, we recorded 1) the number of significant SNP-QT associations linked to the gene, and 2) the smallest p-value among these associations. Similarly, we grouped our QTs into 7 resting state networks (RSNs) using the Yeo atlas, and generated a high-level gene-RSN association map. For each gene-RSN pair, we recorded 1) the number of significant SNP-QT associations linked to the corresponding gene and RSN, and 2) the smallest p-value among these associations.
Results: Using $p \leq 0.05$, our fine-level SNP-QT map included 95,341 SNPs and 68 QTs. The most significant association is between TREM2-rs3095327 and lh_thickness_insula (p=5.1e-10). Our intermediate-level gene-QT map included 19 genes and 68 QTs. The gene containing the most SNP-QT associations was TREM2 (n=282,788), and the top QT was lh_thickness_insula (n=17,032). Hierarchical clustering of our gene-QT -log(p) map identified a subset of genes {TREM2, SPI1, BIN1, APP, PLCG2, CR1, ADAM10} associated with most of the QTs. Our high-level gene-RSN map included 19 genes and 7 RSNs. Default Mode Network was the top RSN with most SNP-QT associations for all 19 genes. TREM2, APP and PLCG2 were the three top genes with most SNP-QT associations for all 7 RSNs.
Conclusions: We reported top findings in mapping associations between cortical thickness QTs and SNPs from AD genes at SNP-QT, gene-QT and gene-RSN levels. The resulting bipartite graphs between these genomic and neuro-phenotypic entities provide valuable information calling for further investigation to guide subsequent studies on disease modeling and drug target discovery.

Keywords: Brain imaging genetics, knowledge graph, cortical thickness, neurodegeneration, Alzheimer's disease

---

Concurrent Session – Application of machine learning techniques in genetics and genomics
Tuesday, August 9, 2022
9:30 AM - 12:05 PM
Grand Ballroom

Pathogenicity prediction for nonsynonymous SNVs and non-frameshift Indels

Xiaoming Liu, The University of South Florida

Multiple computational approaches have been developed to improve our understanding of genetic variants; however, their ability to identify rare pathogenic variants from rare benign variants is still lacking. Using context annotations and deep learning methods, we present pathogenicity prediction models, MetaRNN and MetaRNN-indel, to help identify and prioritize rare nonsynonymous single nucleotide variants (nsSNVs) and non-frameshift insertions/deletions (nfINDELs). We use independent test datasets to demonstrate that these new models outperform state-of-the-art competitors and achieve a more interpretable score distribution. Importantly, prediction scores from both models are comparable, enabling easy adoption of integrated genotype-phenotype association analysis methods.

Transformer-based unsupervised learning for spatial transcriptomic analysis

Chongyue Zhao, University of Pittsburgh

Existing reference-based deconvolution methods integrate single-cell reference and spatial transcriptomics data to predict the proportion of cell-types, but the availability of suitable single-cell reference is often limited. In this talk, we propose a novel Transformer based model to integrate the spatial gene expression measurements and their spatial patterns in the histology image without single cell reference. Our method enables the learning of the locally realistic and globally consistent constituents at nearly single cell resolution.

Assessing Tissue-specific Functional Effects of Non-coding Variants with Deep Learning

Bingshan Li, Vanderbilt University

Abstract: Analysis of whole-genome sequencing (WGS) for genetics is still a challenge due to the lack of accurate functional annotation of noncoding variants, especially the rare ones. As eQTLs have been extensively implicated in the genetics of human diseases, we hypothesize that rare noncoding variants discovered in WGS play a regulatory role in predisposing disease risk. We developed a multi-label learning-based deep neural network to predict the functionality of noncoding variants in the genome based

on eQTLs across 49 human tissues in the GTEx project. TVAR learns the relationships between high-dimensional epigenomics and eQTLs across tissues, taking the correlation among tissues into account to understand shared and tissue-specific eQTL effects. We evaluate TVAR's performance on four complex diseases (coronary artery disease, breast cancer, Type 2 diabetes, and Schizophrenia), using TVAR's tissue-specific annotations, and observe its superior performance in predicting functional variants for both common and rare variants, compared to five existing state-of-the-art tools. We further evaluate TVAR's G-score, a scoring scheme across all tissues, on ClinVar, fine-mapped GWAS loci, Massive Parallel Reporter Assay (MPRA) validated variants, and observe the consistently better performance of TVAR compared to other competing tools.

CNN Algorithms for Disease Classification and Biomarker Discovery

Xiangning Chen, UTHealth Science Center

Convolutional neural network (CNN) has been used broadly in image classification and computer vision, and has achieved high accuracy and reliability. But its application in analyses of genetic and genomic data is limited. In this presentation, I will introduce a technique that transforms genetic and genomic data into artificial image objects (AIO), that in turn could be analyzed effectively with CNN algorithms. The concept is that we consider a variable in a dataset as image pixel, this allows us to utilize the values observed in a given individual to create an AIO. I will present the results obtained from this technique when it was applied to disease classification and biomarker discovery. In a study with 4,096 GWAS p-value selected SNPs, we created 64×64 AIOs for each individual in the Molecular Study of Schizophrenia (MGS), Swedish Case Control Study of Schizophrenia (SCCSS) and the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) datasets, and obtained a classification accuracy of 0.678 ± 0.007, and 0.738 ± 0.008, better than the results obtained from random forest and SVM approaches using the same datasets. In another study, we applied this technique to RNA sequencing datasets (GSE81538 and GSE96058) to classify breast cancer biomarkers, and obtained classification of accuracy of 0.821 ± 0.023 for KI67 and 0.820 ± 0.012, for Nottingham Histologic Grade. These results were better than the multigene classifiers used in the original study. In the third study, we applied this technique to subtype triple negative breast cancer (TNBC), and we identified 21 genes that could divide TNBCs into two major groups with differential overall survival rate ($P = 0.0074$).

Genome-wide cell-free DNA fragmentation as a biomarker for early detection of cancer

Stephen Cristiano, Johns Hopkins University

The high morbidity and mortality of cancer results from late diagnosis where therapeutic intervention is less effective, yet clinically proven biomarkers to broadly diagnose patients are not widely available. Analyses of cell-free DNA (cfDNA) in blood provide a noninvasive diagnostic avenue for patients with cancer. However, cfDNA analyses have largely focused on targeted sequencing of specific genes. Genome-wide analyses of cfDNA features may increase the resolution of changes in circulating tumor

DNA compared to healthy cfDNA and promote more sensitive cancer detection. We developed an approach to analyze fragmentation profiles and cfDNA features across the genome and applied this method to analyze cfDNA from 236 patients with breast, colorectal, lung, ovarian, pancreatic, gastric, or bile duct cancers and 245 healthy individuals. Machine learning incorporating these features resulted in sensitivities of detection from 57% to >99% among seven cancer types at 98% specificity, as well as narrowed the tissue of origin to a limited number of sites. The results of these analyses highlight important properties of cfDNA and provide a facile approach for early detection of human cancer.

Deep learning predicts DNA methylation regulatory variants in the human brain and elucidates the genetics of psychiatric disorders

Jiyun Zhou[1], Qiang Chen[1], Patricia Braun[2], Kira Mandell[2], Andrew Jaffe[1], Haoyang Tan[1], Thomas Hyde[1], Joel Kleinman[1], James Potash[2], Gen Shinozaki[3], Daniel Weinberger[1], Shizhong Han[1*]

[1]Lieber Institute for Brain Development, MD USA
[2]Johns Hopkins University, MD USA
[3]Stanford University, CA USA

There is growing evidence for the role of DNA methylation (DNAm) quantitative trait loci (mQTLs) in the genetics of complex traits, including psychiatric disorders. However, due to extensive linkage disequilibrium (LD) of the genome, it is challenging to identify causal genetic variations that drive DNAm levels by population-based genetic association studies. This limits the utility of mQTLs for fine-mapping risk loci underlying psychiatric disorders identified by genome-wide association studies (GWAS). Here we present INTERACT, a novel deep learning model that integrates convolutional neural network (CNN) with transformer, to predict effects of genetic variations on DNAm levels at CpG sites in the human brain. We show that INTERACT-derived DNAm regulatory variants are not confounded by LD, are concentrated in regulatory genomic regions in the human brain, and are convergent with mQTL evidence from genetic association analysis. We further demonstrate that predicted DNAm regulatory variants are enriched for heritability of brain-related traits and improve polygenic risk prediction for schizophrenia across diverse ancestry samples. Finally, we applied predicted DNAm regulatory variants for fine-mapping schizophrenia GWAS risk loci and identify potential novel risk genes. Our study shows the power of a deep learning approach to identify functional regulatory variants that may elucidate the genetic basis of complex traits.

Machine learning approaches to enhance gene expression prediction integrating eQTLs with 3D genomes and epigenetic data

Dajiang Liu, Ph.D., The Pennsylvania State University

Transcriptome-wide association study (TWAS) is a popular approach to link regulatory variants to target genes. It first builds gene expression prediction models from datasets that measure both genotypes and gene expressions. Based on the prediction models, it then imputes gene expression and tests for

associations with the phenotype of interest. Here, we propose an integrative method PUMICE (Prediction Using Models Informed by Chromatin conformations and Epigenomics) to integrate 3D genomic and epigenomic data with expression quantitative trait loci (eQTL) to more accurately predict gene expressions. PUMICE helps define and prioritize regions that harbor cis-regulatory variants, which outperforms competing methods. We further describe an extension to our method PUMICE +, which jointly combines TWAS results from single- and multi-tissue models. Across 79 traits, PUMICE+ identifies 22% more independent novel genes and increases median chi-square statistics values at known loci by 35% compared to the second-best method, as well as achieves the narrowest credible interval size. Lastly, we perform computational drug repurposing and confirm that PUMICE+ outperforms other TWAS methods.

NSP1 inhibition effects on Human 40s ribosomes compared to intermediate carrier animals

Afagh Bapirzadeh[1,2], Mitra Salehi[1], Zarrin Minuchehr[2], Bijan Bambai[2]

[1]Islamic Azad University, Tehran branch, Heravy Sq., South makran St.,P.O. BOX:1651153311
[2]National Institute of Genetic Engineering and Biotechnology, Pajohesh Blvd., Tehran-Karaj Highway,Tehran, Iran, P.O. Box: 1497716316

Afagh.bapirzadeh@gmail.com,Mitra_salehi_microbiology@yahoo.com,
Zarrin.minuchehr@gmail.com, Bambai2biotech@gmail.com

Abstract
NSP1 protein is one of the first translated proteins of SARS-CoV-2, the virus which is responsible for Corona disease, located at the beginning of 5′end of virus mRNA gene number one. NSP1 binding to Human 40S ribosomal subunits has potential roles in host cells mRNA translation inhibition. To clarify subtle molecular reasons of carrier animal's lack of pathogenicity, we used bioinformatics tools to analyze 7k5i structure from PDB data bank to investigate interactions and internal bonds of 40S ribosome components in normal cell conditions. NSP1 protein bonds in complex with ribosomes of Human host cells in individuals affected by SARS-CoV-2 differ from normal states. Majority of NSP1 interactions involve 18SrRNA ribosomal subunit and one of ribosomal proteins. These bonds are at A605and G600,601 nucleotide positions in 18SrRNA leading to both physical and spatial disrupter of existing internal interactions in this subunit which eventually could cause ribosome to lose its function. Additional to 18SrRNA, NSP1 also bonds with three ribosomal proteins: S2, S3 and S30.alignment comparison of these molecules in Human and other mammalian in one hand and between Human and carrier animals such as camel and Manis in another hand showed for example camel S2 protein is more like Manis instead of being genetically close to human or cows. These crucial findings strongly support important role of NSP1 in translation inhibition of host cells ribosomes.by using these results and making a few structural changes at carboxyl end of NSP1 protein suppression and restriction of cancerous cell growth can be gained.

Keywords: SARS-CoV-2, NSP1, Bioinformatics,40s ribosomal subunits, Bats, Manis

---

Prediction of the effects of missense mutations on human Myeloperoxidase protein stability using in silico saturation mutagenesis

Adebiyi Sobitan[1], William Edwards[1], Md Shah Jalal[1], Kolawole Ayanfe[1], Hemayet Ullah[1], Atanu Duttaroy[1], Jiang Li[2], Shaolei Teng[1*]

[1]Department of Biology, Howard University, Washington DC, 20059 USA
[2]Department of Electrical Engineering and Computer Science, Howard University, Washington DC, 20059 USA
* Correspondence: Email: shaolei.teng@howard.edu (ST)

Abstract
Myeloperoxidase (MPO) is a heme peroxidase with microbicidal properties. MPO plays a role in the host's innate immunity by producing reactive oxygen species inside the cell against foreign organisms. However, there is little functional evidence linking missense mutations to human diseases. We utilized in silico saturation mutagenesis to generate and analyze the effects of 10811 potential missense mutations on MPO stability. Our results showed that ~71% of the potential missense mutations destabilize MPO, and ~8% stabilize the MPO protein. We showed that G402W, G402Y, G361W, G402F, and G655Y would have the highest destabilizing effect on MPO. Meanwhile, D264L, G501M, D264H, D264M, and G501L have the highest stabilization effect on the MPO protein. Our computational tool prediction showed the destabilizing effects in 13 out of 14 MPO missense mutations that cause diseases in humans. We also analyzed putative post-translational modification (PTM) sites on the MPO protein and mapped the PTM sites to disease-associated missense mutations for further analysis. Our analysis showed that R327H associated with frontotemporal dementia and R548W causing generalized pustular psoriasis are near these PTM sites. Our results will aid further research into MPO as a biomarker for human complex diseases and a candidate for drug target discovery.

Keywords: Myeloperoxidase (MPO); in silico saturation mutagenesis; missense mutations; post translational modification (PTM) sites; protein stability.

DelInsCaller: An Efficient Algorithm for Identifying Delins from Long-Read Sequencing Data with High-Level of Sequencing Errors

Shenjie Wang[1,2], Xuanping Zhang[1,2], Geng Qiang[1,2], Jiayin Wang[1,2*]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[2]Shaanxi Engineering Research Center of Medical and Health Big Data, Xi'an Jiaotong University, Xi'an 710049, China

*Correspondence: wangjiayin@mail.xjtu.edu.cn

Abstract
Delins, also known as complex indel, is a combined genomic structural variation that formed by simultaneously deleting and inserting DNA fragments of different sizes at a common genomic location. Recent studies emphasized the importance of delins in diagnose and treatment of many kinds of cancer. However, the existing approaches often encounter two computational problems: 1) delins is difficult to accurately detect. In addition, 2) while the PacBio CLR sequencing achieves significantly longer

reads, the high-level of sequencing errors hurt the delins detection. Thus, in this paper, we proposed an efficient algorithm, named delInsCaller, to identify the delins on haplotype resolution from the PacBio CLR sequencing data. DelInsCaller implemented a highly fault-tolerant method based on variation density scores that can accurately identify regions of delins on haplotype resolution from the PacBio CLR sequencing data with high sequencing errors. We conducted a series of experiments on simulated datasets, and the results showed that delInsCaller outperforms to several state-of-the-art approaches, e.g. SVseq3, across a wide range of sequencing parameters, such as sequencing depth, sequencing errors, etc. delInsCaller often obtained higher f-measures than the existing algorithms, specifically, it maintained the advantages at ~15% sequencing errors. And the f-measure value kept more than 70% on average.

Keywords: Sequencing data analysis; variant calling; Delins; complex indel;

---

microRNA and microRNA target variants associated with autism spectrum disorder and related disorders

Anthony Wong[1], Anbo Zhou[1], Xiaolong Cao[1,5], Vaidhyanathan Mahaganapathy[1], Marco Azaro1, Christine Gwin[1], Sherri Wilson[1], Steve Buyske[2], Christopher W. Bartlett[3], Judy F. Flax[1], Linda M. Brzustowicz[1,4], Jinchuan Xing[1,4,*]

[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; Department of Pediatrics, College of Medicine, The Ohio State University, Columbus, Ohio, USA
[3]The Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway NJ, USA
[4]Current address: Division of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, China

*Correspondence: jinchuan.xing@rutgers.edu;

Abstract
Autism spectrum disorder (ASD) is a childhood neurological disorder with a complex and heterogeneous genetic etiology. Post-transcriptional regulation of ASD implicated genes by microRNA (miRNA), a class of small non-coding RNAs, could play a role in affecting broad molecular pathways related to ASD and associated neurodevelopmental disorders. Using whole-genome sequencing, we analyzed 272 samples in 73 families in the New Jersey Language and Autism Genetics Study (NJLAGS) cohort. These families were recruited to have at least one ASD patient and were further assessed for language impairment and/or reading impairment. Families were also assessed for social responsiveness and attention deficit hyperactivity disorder. miRNA variants and variants within the associated 3' untranslated region (3' UTR) of target mRNAs were annotated and categorized by autosomal dominant, autosomal recessive, or de novo inheritance patterns for each phenotype. A total of 5,104 miRNA and 1,181,148 3' UTR variants were identified in the dataset. After applying several

filtering criteria, including population allele frequency, brain expression, miRNA functional regions, and inheritance patterns, we identified high-confidence variants in 5 brain-expressed miRNAs (targeting 326 genes) and 3' UTR target regions of 152 genes. Some genes, such as SCP2 and UCGC, are identified in multiple families and are known to be involved in neurodevelopmental processes. Using Gene Ontology overrepresentation analysis and protein-protein interaction network analysis, we identified additional clusters of genes and pathways that are important for neurodevelopment. The miRNAs and miRNA target genes identified in this study are potentially involved in neurodevelopmental disorders and should be considered for further functional studies.

Keywords: whole-genome sequencing, miRNA, autism spectrum disorder, family cohort, 3' UTR, neurodevelopmental disorder

---

TransModDNA: Transformer-based DNA Methylation detection on ionic signals from Oxford Nanopore sequencing data

Xiuquan Wang[1], Mian Umair Ahsan[2], Yunyun Zhou[2*], Kai Wang[2,3*]

[1]Department of Mathematics and Computer Science, Tougaloo College, Jackson, MS, United States
[2]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[3]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

*Corresponding author: wangk@email.chop.edu; zhouy6@email.chop.edu

Abstract
Nanopore long-read sequencing technology has become more and more popular for DNA methylation detection nowadays, since it overcomes existing technical limitations for DNA methylation detection using short-reads sequencing or array-based technologies. A number of
analytical tools, including Nanopolish, Tombo, and our tool DeepMod, have been developed to detect DNA methylation from nanopore sequencing reads. However, these tools need to be
further improved for DNA methylation detection in computational efficiency, prediction accuracy, and contextual interpretation on complex genomics regions such as repetitive regions, low GC density regions, and more. In this project, we proposed a new deep learning-based transformer method to detect DNA methylation on ionic signals from Oxford Nanopore sequencing data. Compared to traditional CNN or RNN deep-learning method, transformers can be more accurate in DNA methylation detection, because it can understand relationship between sequential DNA that are far from each other and pay more attention to its most important part in signals. Our project provides a new avenue for applying natural language algorithms in genomics field.

Keywords: Nanopore long-read sequencing, Deep learning transformer model, DNA methylation

---

mintRULS: Prediction of miRNA-mRNA target site interactions using regularized least square method

Sushil Shakyawar[1], Siddesh Southekal[2] and Babu Guda[1]
[1]University of Nebraska Medical Center, NE, USA;
[2]Eli Lilly and Company, USA

Abstract
Identification of miRNA-mRNA interactions is critical to understand the new paradigms in gene regulation. Existing methods show suboptimal performance owing to inappropriate feature selection and limited integration of intuitive biological features of both miRNAs and mRNAs. The present regularized least square-based method, mintRULS, employs features of miRNAs and their target sites using pairwise similarity metrics based on free energy, sequence and repeat identities, and target site accessibility to predict miRNA-target site interactions. We hypothesized that miRNAs sharing similar structural and functional features are more likely to target the same mRNA, and conversely, mRNAs with similar features can be targeted by the same miRNA. Our prediction model achieved an impressive AUC of 0.93 and 0.92 in LOOCV and LmiTOCV settings, respectively. In comparison, other popular tools such as miRDB, TargetScan, MBSTAR, RPmirDIP, and STarMir scored AUCs at 0.73, 0.77, 0.55, 0.84, and 0.67, respectively, in LOOCV setting. Similarly, mintRULS outperformed other methods using metrics such as accuracy, sensitivity, specificity, and MCC. Our method also demonstrated high accuracy when validated against experimentally derived data from condition- and cell-specific studies and expression studies of miRNAs and target genes, both in human and mouse. Program code of mintRULS is freely available at https://doi.org/10.5281/zenodo.6360587 to the research community.

Keywords: miRNA-Target site interaction, Least-square regression, Nucleotide sequence feature, Pairwise score

Time-varying gene expression network analysis reveals conserved transition states in hematopoietic differentiation between human and mouse

Shouguo Gao[1], Ye Chen[2], Zhijie Wu[1], Sachiko Kajigaya[1], Xujing Wang[3], Neal S Young[1]

[1]Hematopoiesis and Bone Marrow Failure Laboratory, Hematology Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, United States of America
[2]Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, Arizona 86011, United States of America.
[3]Division of Diabetes, Endocrinology, and Metabolic Diseases (DEM), National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20817, United States of America.

Abstract
Background: Computational analyses of gene networks can elucidate hematopoietic differentiation from single cell gene expression data, but most algorithms generate only a single, static network. Because gene interactions change over time, it is biologically meaningful to examine time-varying structures and capture dynamic, even transient states and cell-cell relationships.
Methods: A transcriptomic atlas of hematopoietic stem and progenitor cells, 15,245 human and 17,560 murine, was used for network analysis. We employed Monocle to examine trajectories of differentiation, and obtained three dominant branches (erythroid/megakaryocytic, myeloid, and lymphoid), arising directly from the hematopoietic stem cells in both human and mouse. A statistical algorithm, loggle, was used to infer time-varying networks and explore changes of differentiation gene networks over time. A range of network analysis tools (igraph, CompNet, influential, and network_energy) were used to examine properties of and genes in the inferred networks.
Results: Shared characteristics of attributes during the evolution of differentiation gene networks showed a "U" shape of network density over time for all three branches, for human and mouse. Differentiation appeared as a continuous process, originating from stem cells, through a brief transition state marked by fewer gene interactions, before stabilizing in a progenitor state. Human and mouse share hub genes in evolutionary networks. Three conserved network modules were identified that are critical in the lineage differentiation from stem cells to erythroid, myeloid, and lymphoid progenitor cells.
Conclusions: Our analysis confirmed conservation of network dynamics in the hematopoietic systems of mouse and human, reflected by shared hub genes and network topological changes in differentiation. Both human and mouse displayed transition states. Conservation of networks is a powerful approach to identify processes with similarity between mouse and human, which has implications for studies employing the mouse as a model organism.

Robust Augmenting Single-cell RNA-seq with Surface Protein Levels using Geneset Deep Learning and Transfer Learning

Md Musaddaqul Hasib[1,3], Tinghe Zhang[1], Jianqiu Zhang[1], Shou-jiang Gao[2,3], and Yufei Huang[3,4,5*]

[1]Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, USA
[2]Department of Microbiology and Molecular Genetics, University of Pittsburgh, Pittsburgh, PA 15232, USA
[3]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15232, USA
[4]Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15232, USA
[5]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15232, USA

*Correspondence: yuh119@pitt.edu

Abstract
As scRNA-seq become increasingly accessible, providing a cost-efficient method to augment surface protein levels from gene expression measurements are desirable. We proposed a machine learning approach that includes a novel geneset neural network (GS-NN) that aims to learn robust and biologically meaningful features and a highly efficient transfer learning strategy to address cross-dataset differences. We conducted comprehensive experiments to show the improvements of the proposed methods. Specifically, we demonstrate that GS-NN learns more robust features to achieve better cross-subject performance than other machine learning approaches. Transfer learning further improves that of GS-NN by reducing dataset differences through highly efficient fine-tuning. The unique genesets design of GS-NN also allows identification of functions contributing to the prediction and improvement of the proposed strategy. Overall, this study reports a novel approach to robustly augment.

Keywords: CITE-seq; Deep Learning; Transfer Learning; Surface Proteins; Single cell RNA-seq.

GenKI: a variational graph autoencoder based virtual knockout tool for gene function predictions via single-cell gene regulatory network

Yongjian Yang[1], Guanxun Li[2], Yan Zhong[3], Qian Xu[4], James J. Cai[1,4,5]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[2]Department of Statistics, Texas A&M University, College Station, TX 77843, USA
[3]Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, School of Statistics, East China Normal University, 3663 North Zhongshan Road, Shanghai, 200062, China

[4]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA
[5]Interdisciplinary Program of Genetics, Texas A&M University, College Station, TX 77843, USA

Abstract
The latest single-cell RNA sequencing (scRNA-seq) technology facilitates the high-throughput study of cell populations and gene function. For example, scRNA-seq of gene knockout (KO) samples is an unprecedently powerful approach to elucidate gene function. However, due to a lack of experimental and animal resources, exhaustive KO experiments targeting a variety of genes of interest are still not feasible. Here, we present GenKI (GENe Knockout Inference), a computational tool based on variational graph autoencoder (VGAE) to perform virtual gene KO, only requiring scRNA-seq data as input, for gene function study. In the GenKI analysis, a single-cell gene regulatory network (scGRN) is first constructed from scRNA-seq data of a wild-type (WT) sample, and a mirrored KO sample is obtained by virtually deleting a target gene from the WT sample, for which we specifically attenuate the target gene expression to zeros and remove all its edges linking to its regulated genes from the constructed GRN. We use the WT sample data to train the VGAE model, which consists of a two-layer graph convolutional network (GCN) encoder and an inner product decoder. After training, we feed the trained model with the virtual KO sample data. In both cases, we collect the parameters (means and covariances) of the latent bivariate Gaussian distribution for each gene to assess the level of perturbation onto genes by the KO. Finally, a permutation test is used to identify significantly perturbed genes, which are then used to infer the functions of the KO gene and biological processes involved in selected cells. We apply GenKI to real scRNA-seq data to study the function of the triggering receptor expressed on myeloid cells 2 (*Trem2*) in microglia. We demonstrate that the GenKI analysis detects 148 genes that are significantly perturbed by the *Trem2*-KO, and the enrichment analysis identifies lipid metabolism as a significantly enriched function, revealing that *Trem2* plays a critical role in regulating lipid metabolism. This finding is consistent with the conclusion of many other Trem2-KO studies. Additionally, we show that *Apoe* and *Lpl* rank at the top among those significant genes, suggesting *Trem2* regulates *Apoe* and *Lpl*, which are involved in lipid transport and catabolism in microglia. In summary, we show GenKI, the graph-based model which leverages scGRNs, can accurately capture node features and recover the KO gene's functions. We believe our model will make KO-based hypothesis generation and experimental design more feasible and efficient.

Keywords: scRNA-seq, gene knockout, gene regulatory network, machine learning, graph neural network, variational graph autoencoder

---

EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps

Xiaotao Wang[1], Yu Luan[1], Feng Yue[1,2]

[1]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine Northwestern University, Chicago, Illinois, USA
[2]Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, Illinois, USA.

Abstract
Hi-C technique has been shown to be a promising method to detect structural variations (SVs) in human genomes. However, algorithms that can use Hi-C data for a full-range SV detection have been severely lacking. Current methods can only identify inter-chromosomal translocations and long-range intra-chromosomal SVs (>1Mb) at less-than-optimal resolution. Therefore, we develop EagleC, a framework that combines deep-learning and ensemble-learning strategies to predict a full-range of SVs at high-resolution. Importantly, we show that EagleC can uniquely capture a set of fusion genes that are missed by WGS or nanopore. Furthermore, EagleC also effectively captures SVs in other chromatin interaction platforms, such as HiChIP, ChIA-PET, and capture Hi-C. We apply EagleC in over 100 cancer cell lines and primary tumors, and identify a valuable set of high-quality SVs. Finally, we demonstrate that EagleC can be applied to single-cell Hi-C and used to study the SV heterogeneity in primary tumors.

Keywords: Cancer genomes, Structural variations, Deep learning, 3D genome organization, Single-cell Hi-C

This work has been published in Science Advances[1]:

1. Wang, X., Luan, Y. & Yue, F. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. Sci Adv 8, eabn9215 (2022).

Estimation of clonal dynamics of lung cancers based on DNA sequencing data: Potential impact on detection and cure

Marek Kimmel[1], Khanh Dinh[2], Andrew Koval[3]

[1]Departments of Statistics and Bioengineering, Rice University, Houston, TX, USA
[2]Department of Statistics and Irving Institute for Cancer Dynamics, Columbia University, New Yoork, NY, USA
[3]Department of Statistics, Rice University, Houston, TX, USA.

Abstract
Lung cancer (LC) is a deadly disease still claiming around 200,000 lives per year in the United States. As demonstrated, by clinical trials and mathematical modeling, early detection by computed tomography reduces mortality by ca. 20%. However, early LCs detected by screening are curable only in about 50% of cases and deciding which will be cured is difficult. This is likely one of the reasons that screening has not become as widespread as it was hoped. However, due to recent progress in DNA sequencing, including single-cell sequencing, insights can be gained into the timing of the waves of mutations and other genome transformations, which leave trace in the cancer cell genomes. Analysis based on probabilistic models of molecular evolution can help estimate the relative rates of evolution of different clones and hence the relative urations of phases corresponding to small vs. large and slowly vs. fast growing tumors. The relevant theory was published by Dinh et al. (Statistical Science, 2020, 35: 129–144). This talk will show how estimation of clonal dynamics of tumor growth works in simulations, and what are the difficulties encountered in analysis of real data. In addition, we will show

dynamical profiles of LC tumors which might be cured if detected early and of those for which this is more difficult.

Keywords: lung cancer screening, early detection, mathematical modeling, DNA sequencing, site-frequency spectrum

---

Combination of serum and plasma biomarkers could improve prediction performance for Alzheimer's Disease

Fan Zhang[1], Melissa Petersen[1], Leigh Johnson[1], James Hall[1] and Sid O'Bryant[1]
[1]University of North Texas Health Science Center at Fort Worth, TX, USA

Abstract
Background: Alzheimer's disease can be predicted by either serum biomarkers or plasma biomarkers. A combination of both serum and plasma biomarkers may increase the predictive power for Alzheimer's disease. However, due to the high complexity of machine learning, it may also incur overfitting problem.
Objective: In this paper, we investigated whether combining serum and plasma biomarkers with feature selection can improve prediction performance for AD.
Methods: 150 AD patients and 150 normal controls were enrolled for serum test, and 100 AD patients and 100 normal controls were enrolled for plasma test. Among them, 79 ADs and 65NCs are in common. 10 times repeated 5-fold cross-validation model and feature selection method were used to overcome the overfitting problem when combining serum biomarkers and plasma biomarkers together to predict AD.
Results: First, we tested if simply adding serum and plasma biomarkers could make prediction performance improved but also cause overfitting problem. Then we employed a feature selection algorithm we developed to overcome the overfitting problem. Lastly, we tested the prediction performance in two models: 1) training only and 2) 10 times repeated 5-fold cross validation.
Conclusion: We found that combination of serum and plasma biomarkers could improve the prediction performance of AD but might also cause overfitting problem. A further feature selection based on the combination of serum and plasma biomarkers could solve the overfitting problem and produce even higher prediction performance than either a combination of only serum biomarkers or a combination of only plasma markers. The combined feature-selected serum-plasma biomarker approach may have critical implications for understanding the pathophysiology of AD and for developing its preventative treatments.

Keywords: Alzheimer’s Disease, Blood Biomarkers, Support Vector Machine, Machine Learning, Feature Selection

---

Concurrent Session – Machine learning in biomedical research
Tuesday, August 9, 2022
1:40 - 3:00 PM
Grand Ballroom

Secure and Efficient Implementation of Facial Emotion Detection for Smart Patient Monitoring System

Kh Shahriya Zaman, Md Mamun Bin Ibne Reaz

Department of Electrical, Electronic and Systems Engineering Universiti Kebangsaan Malaysia (UKM) Bangi, Malaysia

Abstract
Facial emotion recognition (FER) plays a crucial role in understanding the emotion of people, as well as in social interaction. Numerous advances in machine learning have enabled the automatic detection of facial expressions from a given image of the subject's face. This can be particularly beneficial in smart monitoring and understanding the mental state of medical and psychological patients. With the advent of the Internet-of-Things and Big Data analysis, it is possible to create an intelligent ecosystem for continuous patient monitoring with minimal human interaction. Most algorithms that attain high emotion classification accuracy require extensive computational resources, which either require bulky and inefficient devices or require the sensor data to be processed on cloud servers. However, there is always the risk of privacy invasion, data misuse, and data manipulation when the raw images are transferred to cloud servers for processing FER data. One possible solution to this problem is to minimize the movement of such private data. In this research, we propose an efficient implementation of a convolutional neural network (CNN) based algorithm for on-device FER on a low-power FPGA platform. This is done by encoding the CNN weights to approximated signed digits, which reduces the number of partial sums to be computed for multiply-accumulate (MAC) operations. This is advantageous for portable devices that lack full-fledged resource-intensive multipliers. We implemented several CNN models on the FPGA for FER and presented their respective detection accuracy. Our implementations reduce the FPGA resource requirement by at least 22% compared to models with integer weight, with negligible loss in classification accuracy. The outcome of this research will help in the development of secure and low-power systems for FER and other biomedical applications.

Keywords: facial expression detection, emotion recognition, FPGA implementation, convolutional neural network, signed digit approximation

---

DeepG2P: predicting protein abundance from mRNA expression

Hui-Mei Tsai[1,2,3*], Tzu-Hung Hsiao[2*], Yu-Chiao Chiu[3], Yufei Huang[4,5], Yidong Chen[3,6§], Eric

Y. Chuang[1§]

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan
[2]Department of Medical Research, Taichung Veterans General Hospital, Taichung 40705, Taiwan
[3]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[4]Department of Medicine, School of Medicine, University of Pittsburgh, PA 15232, USA
[5]UPMC Hillman Cancer Center, University of Pittsburgh, PA 15232, USA
[6]Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA
*Co-first

§Correspondence: Eric Y. Chuang (chuangey@ntu.edu.tw); Yidong Chen (cheny8@uthscsa.edu)

Abstract
Background: Central dogma is the basic principle of life science, yet its inner, cryptic regulation is a daunting challenge in decoding genome. Central dogma states a two-step process of genetic information flow, that is, DNA undergoes transcription to produce RNA and RNA undergoes translation to produce protein. However, this seemingly straightforward process does not fully reflect the reality as various modifications and regulatory molecules are discovered which increases the complexities of the conventional central dogma. Only one-third of proteins can be predicted from their corresponding RNA expression.
Results: We proposed a deep learning model (DeepG2P) to predict protein abundance based on the genome-wide transcriptome profile of pan-cancers on the Cancer Genome Atlas (TCGA) and the Cancer Proteome Atlas (TCPA). DeepG2P outperformed three conventional machine learning methods with averaged Pearson correlation of 0.656 across 187 proteins. We also showed that our multi-task DeepG2P has equivalent prediction performance to single-task DeepG2P models. To explain our DeepG2P model between input RNA expression (predictor) to predicted protein abundance, we used integrated gradients to extract top predictors and showed that their expression is indeed associated with protein abundance. And as the number of predictors increased, tumor types become apparent.
Conclusions: Here we present a CNN model that can predict multiple cancer-associated proteins simultaneously based on gene expression profiles. The results showed protein abundance prediction improvement. We expect the DeepG2P model will enable novel mechanisms studies at the protein level, in silico.

Keywords: Deep learning, Convolutional neural networks, Protein prediction, Proteome, Transcriptome, The Cancer Genome Atlas, The Cancer Proteome Atlas

---

Transformer for Gene Expression Modeling (T-GEM): An interpretable deep learning model for gene expression-based phenotype predictions

Ting-He Zhang[1], Md Musaddaqui Hasib[1], Yu-Chiao Chiu[2], Zhi-Feng Han[1], Yu-Fang Jin[1], Mario Flores[1], Yidong Chen[2,*], Yufei Huang[3,4,*]

[1]Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, 78249, USA
[2]Greehey Children's Cancer Research Institute, The University of Texas Health Science Center at San Antonio, San Antonio, TX, 78229, USA
[3]Department of Medicine, School of Medicine, University of Pittsburgh, PA 15232, USA
[4]UPMC Hillman Cancer Center, University of Pittsburgh, PA 15232, USA

*Correspondence: Y,H : YUH119@pitt.edu ; Y,C : ChenY8@uthscsa.edu

Abstract
Background: Deep learning has been applied in precision oncology to address a variety of gene expression-based phenotype predictions. However, gene expression data's unique characteristics challenge the computer vision inspired design of popular DL models such as CNN and ask for the need to develop interpretable DL models tailored for transcriptomics study.
Result: To address the current challenges in developing interpretable DL model for modeling gene expression data, we propose a novel interpretable deep learning architecture called T-GEM, or Transformer for Gene Expression Modeling. We provided the detailed T-GEM model for modeling gene-gene interactions and demonstrated its utility for gene expression-based predictions of cancer-related phenotypes including cancer type prediction and immune cell type classification. We carefully analyzed the learning mechanism of T-GEM and showed that the first layer has broader attention while higher layers focus more on phenotype related genes. We also showed that T-GEM's self-attention could capture important biological functions associated with the predicted phenotypes. We further devised a method to extract the regulatory network that T-GEM learns by exploiting the attributions of self-attention weights for classifications and showed that the network hub genes were likely markers for the predicted phenotypes.

Keywords: phenotypes prediction; interpretable deep learning; Transformer; cancer type prediction; immune cell type prediction

---

Interpretable Drug Synergy Prediction with Graph Neural Networks for Human-AI Collaboration in Healthcare

Zehao Dong[1], Heming Zhang[1], Yixin Chen[1], Fuhai Li[2,3]

[1]Computer Science, Washington University in St. Louis, St. Louis, MO, USA
[2]Institute for Informatics (I2), Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA
[3]Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA

zehao.dong@wustl.edu; hemingzhang@wustl.edu; chen@cse.wustl.edu; Fuhai.Li@wustl.edu

Abstract
We investigate molecular mechanisms of resistant or sensitive response of cancer drug combination therapies in an inductive and interpretable manner. Though deep learning algorithms are widely used in the drug synergy prediction problem, it is still an open problem to formulate the prediction model with biological meaning to investigate the mysterious mechanisms of synergy (MoS) for the human-AI collaboration in healthcare systems. To address the challenges, we propose a deep graph neural network, IDSP (Interpretable Deep Signaling Pathways), to incorporate the gene-gene as well as gene-drug regulatory relationships in synergic drug combination predictions. IDSP automatically learns weights of
edges based on the gene and drug node relations, i.e., signaling interactions, by a multi-layer perceptron (MLP) and aggregates information in an inductive manner. The proposed architecture generates interpretable drug synergy prediction by detecting important signaling interactions, and can be implemented when the underlying molecular mechanism encounters unseen genes or signaling pathways. We test IDWSP on signaling networks formulated by genes from 46 core cancer signaling pathways and drug combinations from NCI ALMANAC drug combination screening data. The experimental results demonstrated that 1) IDSP can learn from the underlying molecular mechanism to make prediction without additional drug
chemical information while achieving highly comparable performance with current state-of-art methods; 2) IDSP show superior generality and flexibility to implement the synergy prediction task on both transductive tasks and inductive tasks. 3) IDSP can generate interpretable results by detecting different salient gene patterns (i.e. MoS) for different cell lines.

Concurrent Session – COVID-19 Informatics
Tuesday, August 9, 2022
1:40 - 3:00 PM
Logan

Drug-Target Network Study Reveals the Core Target-protein Interactions of Various COVID-19 Treatments

Yulin Dai [1,‡], Hui Yu [2,‡], Qiheng Yan [1,3], Bingrui Li [1,4], Andi Liu [1,5], Wendao Liu [1,6], Xiaoqian Jiang [7], Yejin Kim [7], Yan Guo [2,*], and Zhongming Zhao [1,5,6,8,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX 77030, USA
[2]Comprehensive Cancer Center, Department of Internal Medicine, The University of New Mexico, Albuquerque, NM 87131, USA
[3]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[4]Metastasis Research Center, Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[5]Department of Epidemiology, Human Genetics and Environmental Sciences, 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
[7]Center for Secure Artificial Intelligence for Healthcare, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[8]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[‡]These authors contribute equally.
[*]Correspondence: Zhongming Zhao: Zhongming.Zhao@uth.tmc.edu, The University of Texas Health Science Center at Houston 7000 Fannin St. Suite 600, Houston, TX 77030; Yan Guo; YaGuo@salud.unm.edu, Comprehensive Cancer Center, Department of Internal Medicine, The University of New Mexico, Albuquerque, NM 87131, USA

Abstract
Coronavirus Disease-2019 (COVID-19) pandemic has caused a dramatic loss of human life and devastated the worldwide economy. Numerous efforts have been spent to mitigate COVID-19 symptoms and reduce the death rate. We conducted literature mining for more than 250 thousand published works and curated 174 most widely used COVID-19 medications. Overlaid with human protein-protein interaction (PPI) network, we used Steiner tree analysis to extract a core subnetwork that grew from the pharmacological targets of ten credible drugs ascertained by CTDbase. The resultant core subnetwork consisted of 34 interconnected genes, which were associated with 36 drugs. Immune cell membrane receptors, downstream cellular signaling cascade, and severe COVID-19 symptom risk

were significantly enriched for the core subnetwork genes. The lung mast cell was most enriched for target genes among 1,355 human tissue-cell types. Human bronchoalveolar lavage fluid COVID-19 single-cell RNA-Seq data highlighted T cells and macrophages have the most overlapping genes from the core subnetwork. Overall, we constructed an actionable human target-protein module that were mainly involved anti-inflammatory/anti-viral entry functions and highly overlapped with COVID-19 severity-related genes. Our findings could serve as a knowledge base for guiding drug discovery or repurposing to confront the fast-evolving SARS-CoV-2 virus and other severe infectious diseases.

Integrated analysis of bulk RNA-seq and single-cell RNA-seq unravels the influences of SARS-CoV-2 infections to cancer patients

Yu Chen[1,2], Yujia Qin[1], Yuanyuan Fu[1], Zitong Gao[1,2], Youping Deng[1,*]

[1]Department of Quantitative Health Sciences, John A. Burns School of Medicine, University of Hawaii at Manoa, 651 Ilalo Street, Honolulu, HI, USA.
[2]Department of Molecular Biosciences and Bioengineering, College of Tropical Agriculture and Human Resources, University of Hawaii at Manoa, 1955 East West Road, Agricultural Sciences, Honolulu, HI, USA.

*Correspondence
Youping Deng, Department of Quantitative Health Sciences, John A. Burns School of Medicine, University of Hawaii at Manoa, 651 Ilalo Street, Honolulu, HI, USA. Email: dengy@hawaii.edu

Abstract
Syndromic coronavirus 2 (SARSCoV2) is a highly contagious and pathogenic coronavirus that emerged in late 2019 and caused a pandemic of respiratory illness termed as coronavirus disease 2019 (COVID19). Cancer patients are more susceptible to SARS-CoV-2 infection. The treatment of cancer patients infected with SARSCoV2 is more complicated and the patients are at risk of poor prognosis compared to other populations. Patients infected with SARS-CoV-2 are prone to rapid development of acute respiratory distress syndrome (ARDS), of which, pulmonary fibrosis (PF) is considered as a sequelae. Both ARDS and PF are factors that contribute to poor prognosis in COVID-19 patients. However, the molecular mechanisms among COVID-19, ARDS and PF in COVID-19 patients with cancer are not well understood. In this study, the common differentially expressed genes (DEGs) between COVID19 patients with and without cancer were identified. Based on the common DEGs, a series of analyses were performed, including Gene ontology (GO) and pathway analysis, protein-protein Interaction (PPI) network construction and hub gene extraction, transcription factor (TF)-DEG regulatory network construction, TF-DEG-miRNA coregulatory network construction and drug molecule identification. The candidate drug molecules (e.g. Tamibarotene CTD 00002527) obtained by this study might be helpful for effective therapeutic in COVID-19 patients with cancer. In addition, the common DEGs among ARDS, PF and COVID19 patients with and without cancer are TNFSF10 and IFITM2. These two genes may serve as potential therapeutic targets in the treatment of COVID-19 patients with cancer. Changes in the expression levels of TNFSF10 and IFITM2 in CD14+ and CD16+ monocytes may affect the immune response of COVID-19 patients. Targeting m^6A pathways (e.g. METTL3/SERPINA1 axis) to restrict SARS-CoV-2 reproduction has therapeutic potential for

COVID-19 patients. Importantly, innate and adaptive immune responses was activated after SARS-CoV-2 infection in COVID-19 patient without cancer. However, the suppression of immune responses was more pronounced in COVID-19 patients with cancer.
Keywords: SARS-CoV-2; cancer; pulmonary fibrosis; acute respiratory distress; protein–protein interaction (PPI); drug molecule; single-cell RNA-seq; Immunity; monocyte; m^6A

Cell-Specific Gene Signature for COVID-19 Infection Severity Using single-cell RNA-seq analysis

Mario Flores[1§], Karla Paniagua[1], Wenjian Huang[1], Ricardo Ramirez[2], Yidong Chen[3,4], and Yufei Huang[5,6], Yu-Fang Jin[1§]

[1]Department of Electrical and Computer Engineering, the University of Texas at San Antonio, San Antonio, TX 78249, USA
[2]Department of Electrical Engineering and Cyber Engineering, the Houston Baptist University, Houston, TX77074, USA
[3]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[4]Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[5]Department of Medicine, School of Medicine, University of Pittsburgh, PA 15232, USA
[6]UPMC Hillman Cancer Center, University of Pittsburgh, PA 15232, USA

§Correspondence should be addressed to Yu-Fang Jin (yufang.jin@utsa.edu); and Mario Flores (mario.flores@utsa.edu);

Abstract
SARS-CoV2, the etiological agent responsible for COVID-19 has affected the lives of billions and killed millions of infected people. This virus has been demonstrated to have different outcomes among individuals, some of them presenting a mild infection, while others presenting severe symptoms or even death. The identification of the molecular states related to the severity of COVID-19 infection has become of utmost importance to understanding the critical immune response differences. In this study, we computationally processed public scRNA-Seq data of 12 patients diagnosed as having mild, severe, or no infection and generate a high-quality integrated dataset that consist of 63,734 cells X 23,916 genes. We found significant differences in cell-type composition in mild and severe groups compared to the normal group. Overall, inflammatory responses were dramatically elevated in the severe group, which was evidenced by the significant increase of macrophages from 7.51% in the mild group to 24.29% in the severe group. As an indicator of immune defense, populations of T cells counted for 21.12% in the mild group and decreased by half to 11.15% in the severe group. To verify these findings we developed several artificial neural networks (ANN) and graph convolutional neural networks (GCNN) and show that the GCNN model reach a prediction accuracy of the infection severity of 94.57% using data from subtypes of macrophages. Overall our study indicates significant differences in the gene expression profiles of subtypes of immune cells of severely infected patients.

Classifying Refugee Status Using Common Features in EMR

Malia Morrison[1], Vanessa Nobles[1], Crista E. Johnson-Agbakwu[2,3,4,5], Celeste Bailey[2,3], Li Liu[1,6,7,*]

[1]College of Health Solutions, Arizona State University, Phoenix, AZ, 85004, USA
[2]Department of Obstetrics, Gynecology and Women's Health, Valleywise Health, Phoenix, AZ, 85008, USA
[3]Creighton University School of Medicine -Phoenix Campus, Phoenix, AZ, 85008, USA
[4]District Medical Group, Mesa, AZ, 85201, USA
[5]Southwest Interdisciplinary Research Center, Watts College of Public Service and Community Solutions, Arizona State University, Tempe, AZ
[6]Biodesign Institute, Arizona State University, Tempe, AZ, 85281, USA
[7]Department of Neurology, Mayo Clinic, Scottsdale, AZ 85259, USA

*Corresponding author: Li Liu liliu@asu.edu

Abstract
Objective: Automated and accurate identification of refugees in healthcare databases is a critical first step to investigate healthcare needs of this vulnerable population and improve health disparities. This study developed a machine-learning method, named refugee identification system (RIS) that uses features commonly collected in healthcare databases to classify refugees and non-refugees.
Materials and Methods: We compiled a curated data set consisting of 103 refugees and 930 non-refugees in Arizona. For each person in the curated data set, we collected age, primary language, and noise-masked home address. We supplemented de-identified individual-level data with state-level refugee resettlement statistics and world language statistics, then performed feature engineering to convert primary language and masked address into quantitative features. Finally, we built a random forest model to classify refugee status.
Results: Evaluated on holdout testing data, RIS achieved a high classification accuracy of 0.97, specificity of 0.99, sensitivity of 0.85, positive predictive value of 0.88, and negative predictive value of 0.98. The receiver operating characteristic curve had an area under the curve value of 0.98.
Discussion and Conclusion: RIS is an automated, accurate, generalizable, and scalable method to predict refugee status. It uses only de-identified information to protect patient privacy. RIS enables large-scale investigation of refugee healthcare needs and improvement of health disparities.

Keywords: refugee health, machine learning, health disparity

Using machine learning to assess the range of services provided by family physicians in Ontario, Canada

Arunim Garg[1], David W. Savage[2], Salimur Choudhury[1], and Vijay Mago[1]

[1]Department of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1 Canada
[2]NOSM University, Thunder Bay, ON P7B 5E1 Canada

Abstract
Health human resource planning is about having the right number of physicians available at any given time with the right training. It is considered to be a significant challenge, especially in rural communities in the Northern Ontario. Previous studies have highlighted that the family physicians in rural communities work to a fuller extent of their scope of practice than urban physicians to meet the needs of rural populations. These rural physicians provide essential care that reduces the gap between rural and urban healthcare systems. In Ontario, although the location of practice (north or south) or rurality (urban or rural) of physicians can be determined, the features that define a rural family physician are less well understood. Given this need, the purpose of this research is to determine the factors that predict a physicians rurality and location of practice in Ontario. Physician demographic and practice pattern data were used to predict the four classes for each physician: north-rural, north-urban, south-rural, and south-urban. Machine learning techniques, such as k-nearest neighbours, random forests and gradient boosting machines, are used to solve the classification problem. The best model developed is used to gain meaningful insights into the factors that impact a physician's rurality the most. Additionally, hypothesis testing was then conducted to determine the statistical evidence to validate the results obtained from machine learning techniques. These insights will improve decision-making by the government and other policy makers and potentially improve the health of people in northern and rural communities of Ontario.

TermViewer - A Web Application for Streamlined Human Phenotype Ontology (HPO) Tagging and Document Annotation

Anna Nixon[1], James M. Havrilla[1], Li Fang[1], Kai Wang[1,2]

[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
[2]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

Correspondence: wangk@chop.edu

Abstract
Clinical notes from electronic health records (EHRs) contain a large amount of clinical phenotype data on patients that can provide insights into the phenotypic presentation of various diseases. A number of Natural Language Processing (NLP) algorithms have been utilized in the past few years to annotate medical concepts, such as Human Phenotype Ontology (HPO) terms, from clinical notes. However,

efficient use of NLP algorithms require the use of high-quality clinical notes with phenotype descriptions, and erroneous annotations often exist in results from these NLP algorithms. Manual review by human experts is often needed to compile the correct phenotype information on individual patients. Here we develop TermViewer, a web application that allows multi-party collaborative annotation and quality assessment of clinical notes that have already been processed and tagged by Natural Language Processing (NLP) algorithms. TermViewer allows users to view clinical notes with Human Phenotype Ontology (HPO) terms highlighted, and to easily classify high-quality notes and revise incorrect tagging of HPO terms. Currently, TermViewer combines MetaMap and cTakes, two of the most widely used NLP tools for tagging medical terms, and identifies where these two tools agree and disagree, allowing users to perform manual review of computationally generated HPO annotations. TermViewer can be a stand-alone tool for analyzing notes or become part of a machine learning pipeline where tagged HPO terms can be used as additional input data. TermViewer is available at https://github.com/WGLab/TermViewer.

Mining High-level Imaging Genetic Associations via Clustering AD Candidate Variants with Similar Brain Association Patterns

Ruiming Wu[1], Jingxuan Bao[1], Mansu Kim[1], Andrew Saykin[2], Jason Moore[3] and Li Shen[1]
[1]University of Pennsylvania, PA, USA;
[2]Indiana University, IN, USA;
[3]Cedars-Sinai, CA, USA;

Abstract
Brain imaging genetics examines associations between imaging quantitative traits (QTs) and genetic factors such as single nucleotide polymorphisms (SNPs) to provide important insights into the pathogenesis of Alzheimer's disease (AD). The individual level SNP-QT signals are high dimensional and typically have small effect sizes, making them hard to be detected and replicated. To overcome this limitation, this work proposes a new approach that identifies high-level imaging genetic associations through applying multigraph clustering to the SNP-QT association maps. Given a SNP set and a brain QT set, the association between each SNP and each QT is evaluated using a linear regression model. Based on the resulting SNP-QT association map, five SNP-SNP similarity networks (or graphs) are created using five different scoring functions respectively. Multigraph clustering is apply to these networks to identify SNP clusters with similar association patterns with all the brain QTs. After that, functional annotation is performed for each identified SNP cluster and its corresponding brain association pattern. We applied this pipeline to an AD imaging genetic study, which yielded promising results. For example, in an association study between 54 AD SNPs and 116 amyloid QTs, we identified two SNP clusters with one responsible for amyloid beta clearances and the other regulating amyloid beta formation. These high-level findings have the potential to provide valuable insights into relevant genetic pathways and brain circuits, which can help form new hypotheses for more detailed imaging and genetics studies in independent cohorts.

Keywords: brain imaging genetics, multigraph clustering, Alzheimer's disease

Flash Talk Session I
Monday, August 8, 2022
11:45 AM - 12:15 PM
Logan

Structure-enhanced Deep Meta-learning Predicts Uncharted Chemical-Protein Interactions on a Genome-scale

Tian Cai[1], Li Xie[2], Shuo Zhang[1], Muge Chen[3], Di He[1], Yang Liu[2], Hari Krishna Namballa[4], Michael Dorogan[4], Wayne W. Harding[4], Cameron Mura[5], Philip E.Bourne[5], Lei Xie[1,2,6]

[1]Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, New York, 10016, USA
[2]Department of Computer Science, Hunter College, The City University of New York, New York, 10065, USA
[3]Master Program in Computer Science, Courant Institute of Mathematical Sciences, New York University
[4]Department of Chemistry, Hunter College, The City University of New York, New York, 10065, USA
[5]School of Data Science & Department of Biomedical Engineering, University of Virginia, Virginia, 22903, USA
[6]Helen and Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, 10021, USA

Abstract
Discovering genome-wide chemical-protein interactions is instrumental for chemical genomics, drug discovery and precision medicine. However, more than 90% of gene families remain dark, i.e., their small molecular ligands are undiscovered. Existing approaches typically fail when the dark protein of interest differs from those with known ligands or structures. To address this challenge, we developed a deep learning framework PortalCG. PortalCG consists of three novel components: (i) end-to-end step-wise transfer learning in recognition of sequence-structure-function paradigm, (ii) out-of-cluster meta-learning in light of protein evolution for generalizing machine learning models to unstudied gene families, and (iii) stress model selection to facilitate model deployment in a real-world scenario. In rigorous benchmark experiments, PortalCG considerably outperformed state-of-the-art techniques when applied to dark gene families. Experimental validations on 65 compounds supported the accuracy and robustness of PortalCG. Thus, PortalCG is a viable solution to the out-of-distribution (OOD) problem in exploring the dark protein functional space, and can be applied to a wide variety of scientific domains.

Keywords: deep learning, meta-learning, chemical-protein interaction, dark proteins, drug discovery

Deep Learning Prediction of Chemical-induced Dose-Dependent and Context-Specific Multiplex Phenotype Responses and Its Application to Personalized Alzheimer's Disease Drug Repurposing

You Wu[1], Qiao Liu[1], Yue Qiu[1], Lei Xie[2]

[1]CUNY Graduate Center, NY United States
[2]CUNY Hunter College, NY United States

Abstract
Predictive modeling of drug-induced gene expressions is a powerful tool for phenotype-based compound screening and drug repurposing. State-of-the-art machine learning methods use a small number of fixed cell lines as a surrogate for predicting actual expressions in a new cell type or tissue, although it is well known that drug responses depend on a cellular context. Thus, the existing approach has limitations when applied to personalized medicine, especially for many understudied diseases whose molecular profiles are dramatically different from those characterized in the training data. Besides the gene expression, dose-dependent cell viability is another important phenotype readout and is more informative than conventional summary statistics (e.g., IC50) for characterizing clinical drug efficacy and toxicity. However, few computational methods can reliably predict the dose dependent cell viability. To address the challenges mentioned above, we designed a new deep learning model, MultiDCP, to predict cellular context-dependent gene expressions and cell viability on a specific dosage. The novelties of MultiDCP include a knowledge-driven gene expression profile transformer that enables context-specific phenotypic response predictions of novel cells or tissues, integration of multiple diverse labeled and unlabeled omics data, the joint training of the multiple prediction tasks, and a teacher-student training procedure that allows us to utilize unreliable data effectively.
Comprehensive benchmark studies suggest that MultiDCP outperforms state-of-the-art methods in the setting of novel cell lines. The predicted drug-induced gene expressions demonstrate a stronger predictive power than noisy experimental data for downstream tasks. We applied MultiDCP to repurpose individualized drugs for Alzheimer's disease in terms of efficacy and toxicity, suggesting that MultiDCP is a potentially powerful tool for personalized drug discovery.

Keywords: Deep learning; Drug repurposing; Drug discovery; Computational biology; Alzheimer's disease; Artificial intelligence; Transfer learning

---

SPCount: Advances in deep taxonomy alignments for extracellular small RNAs

Quanhu Sheng[1], Marisol Ramirez[1], Ryan M. Allen[2], Qi Liu[1], Michelle J. Ormseth[2], Kasey C. Vickers[2*], Shyr Yu[1*]

[1]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA
[2]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Extracellular small RNAs (sRNA) are transported in biofluids by lipoproteins. We have previously reported that mammalian lipoproteins are highly-enriched with microbial sRNAs likely derived from bacteria and fungi in the microbiome and environment, including a large number of rRNA-derived sRNAs from Proteobacteria. To quantify non-host sRNAs in lipoprotein datasets, we previously released the TIGER sRNA sequencing data analysis package; however, this software utilized representative genomes for microbial alignment for species summary counts and thus, was limited in the depth of taxonomical coverage. To identify more microbial sRNAs through increased genome alignments and overcome current limitations in analysis, we developed a novel python package entitled "Small RNA Perfect-match Counting (SPCount)". SPCount contains three major functions: database preparation, read mapping and counting at sample level, and count table assembly at project level. Based on user-provided root taxonomy id, SPCount will download the genome sequences of species belonging to the root taxonomy entry and build corresponding genome databases ready for mapping. SPCount generates mapping scripts for each sample using bowtie and those scripts can be processed through local computer or by cluster. SPCount outputs include query count, fraction-based estimated count, and unique-mapped count at six taxonomy levels, including species, genus, family, order, class and phylum. In addition, SPCount also outputs a read sequence-based count table and uniquely mapped count table on multi-ranks. To compare SPCount to the initial TIGER pipeline, we tested the impact of the environment on extracellular sRNAs circulating on mouse high-density lipoproteins (HDL). Mice were housed in either a sterile germ-free room or a control specific pathogen-free (SPF) room for 30 days. HDL were isolated from plasma by size-exclusion chromatography and sRNA sequencing was completed on HDL samples using degenerate-base adapter libraries and Illumina sequencing. Remarkably, SPCount identified on 1,048,575 unique reads and 301,479.6 total Reads Per Million (RPM, mean per sample) compared to TIGER identification of 395,997 unique reads and 169,442 RPM total reads, a 2.64-fold and 1.78-fold increase in data, respectively. DEseq2 was used to identify significant differentially altered bacterial sRNAs based on species counts, and SPCount identified 779 significantly altered species compared to only 22 for TIGER, a 35.41-fold increase. Since SPCount is taxonomy based, it can be used not only on bacteria genomes, but also on virus, fungi and other kingdoms. Collectively, SPCount facilitated advanced discovery of many novel bacterial sRNAs on HDL and has tremendous applicability for the field of extracellular sRNA.

Genome-wide analysis on model derived binge-eating disorder phenotype identifies the first three risk loci and implicates iron metabolism

David Burstein[1-7*], Trevor Griffen[1,8*], Karen Therrien[1-7], Jaroslav Bendl[1-6], Sanan Venkatesh[1-7], Pengfei Dong[1-6], Amirhossein Modabbernia[1], Biao Zeng[1-6], Deepika Mathur[1-6], Gabriel Hoffman[1-6], Robyn Sysko[1,8], Tom Hildebrandt[1,8], Georgios Voloudakis[1-7†], Panos Roussos[1-7,9†]

*These authors contributed equally
†These authors jointly supervised this work

[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]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6]Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[7]Mental Illness Research, Education, and Clinical Center (VISN 2 South), James J. Peters VA Medical Center, Bronx, NY, USA
[8]Center of Excellence in Eating and Weight Disorders, Icahn School of Medicine at Mount Sinai, New York, NY USA
[9]Center for Dementia Research, Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA.

Abstract
Binge-eating disorder is a common and heritable psychiatric illness; however, debate about whether and how binge-eating disorder is distinct from obesity led to its exclusion from *The Diagnostic and Statistical Manual of Mental Disorders* (*DSM)* until the publication of *its 5th version in 2013*. We provide, for the first time, a genome-wide association study of binge-eating disorder, implicating three genes and iron metabolism in its pathophysiology.
We leverage machine learning techniques on a very large and diverse cohort, the Million Veteran Program (n = 767,527), to overcome the limited sample size of diagnosed individuals with binge-eating disorder (n = 822) due to its relatively recent inclusion in diagnostic classification systems. We tested our penalized logistic regression model against a holdout group comprising 10% of the cohort and achieved a strong predictive performance: the sensitivity-specificity area under the curve was 97.1%, and, while our prevalence of identified BED cases was only 0.1%, the average positive predictive value was 11.0% Subsequently, we perform a genome-wide association study with our model derived phenotype on individuals of African (*n* = 77,574) and European (*n* = 285,138) ancestry while controlling for body mass index to identify three independent loci near the *HFE*, *MCHR2* and *LRP11* genes.
We then validate our machine learning-based approach by performing LD-score regression to confirm a high genetic correlation (*rg* = 0.85) with our more traditional case-control GWAS. Furthermore, we note that our machine learning approach outperforms the case-control GWAS on both SNP heritability ($p = 6.74\times10\text{-}21$ vs. $p = 0.11$) and polygenic risk score validation meta-analyzed across three external cohorts ($p = 1.39\times10\text{-}3$ vs. $p = 0.44$). Finally, we identify genetic association between BED and several neuropsychiatric traits and implicate iron metabolism in the pathophysiology of BED. Overall, our findings provide insights into the genetics underlying BED and suggest directions for future translational research.

Link to preprint on medRxiv: https://doi.org/10.1101/2022.04.28.22274437

Keywords: EHR-based phenotyping, machine learning, GWAS, eating disorders, binge-eating disorder

A new player into the cellular heterogeneity: the flexible and mobile extrachromosomal circular DNA

Jiajinlong Kang[1], Yulin Dai[1], Jinze Li[2], Huihui Fan[1] & Zhongming Zhao[1,3,4]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA;
[2]Environmental and Occupational Health Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA;
[3]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[4]MD Anderson Cancer Center, University of Texas Health Graduate School of Biomedical Sciences, Houston, TX 77030, USA

Abstract
Extrachromosomal circular DNA (eccDNA) is a special class of DNA derived from linear chromosomes yet existing independently. It has been identified in multiple organisms, including homo sapiens, and has been shown to play important roles relevant to tumor progression and drug resistance. However, previous eccDNA studies and computational tools developed for eccDNA detection are based on bulk samples, therefore lacking sufficient power to reveal the heterogeneous and cell-type-specific landscape of eccDNA in tumor cell and its associated microenvironment. Here, we proposed a novel approach to study eccDNA at the single-cell level via integrating adult and pediatric glioblastoma (GBM) samples profiled using single-cell Assay for Transposase-Accessible Chromatin with sequencing. We provided an overview of the cellular origins, the aligned linear genomic distribution, as well as the differential regulations between linear and circular genome under disease- and cell-type-specific conditions across the open chromatin regions in GBM. Specifically, we highlighted the potential of eccDNA in carrying regulatory sequences and acting as mobile enhancers that may function in a trans-regulation manner. Taken together, our results expand the current understanding of eccDNA in the context of tumorigenesis through single-cell approach, and emphasize the necessity of further probe into eccDNA research at the single-cell level.

Keywords: Single-cell ATAC sequencing; glioblastoma; eccDNAs; distal regulators; trans-regulation.

Fifty-one novel, replicated loci identified in genome-wide association study of polyunsaturated and monounsaturated fatty acids in 124,024 European individuals

Michael Francis[1], Yitang Sun[2], Huifang Xu[2], J. Thomas Brenna[3,4], Kaixiong Ye[1,2]

[1]Institute of Bioinformatics, University of Georgia, Athens, Georgia, US
[2]Department of Genetics, University of Georgia, Athens, Georgia, US
[3]Division of Nutritional Sciences, Cornell University, Ithaca, NY, US
[4]Dell Pediatric Research Institute and the Depts of Pediatrics, of Nutrition, and of Chemistry University of Texas at Austin, Austin, TX, US

Abstract
Circulating polyunsaturated and monounsaturated fatty acid (PUFA and MUFA) levels, whose imbalances co-occur with human metabolic diseases, have strong heritable components. Genome-wide association studies (GWAS) have identified 37 unique genomic loci related to PUFAs and MUFAs. However, they collectively only explain a small fraction of the phenotypic variance, suggesting more loci are left to be found in large samples. Here, we performed the largest GWAS to date on fourteen PUFA and MUFA phenotypes, measured by nuclear magnetic resonance in plasma. First, we performed a discovery GWAS in the European samples from the UK Biobank (n=101,729). We identified 612 significant loci-phenotype associations (115 unique loci; P < 1.678*10-8). Second, for five phenotypes (omega-3, omega-6, DHA, LA, and MUFAs), we performed a replication analysis in two external European studies: FinMetSeq (n=8,751) and a meta-analysis by Kettunnen et al. (n=3,644-13,544). Third, we performed a meta-analysis across these three studies, yielding 254 significant loci-phenotype associations (109 unique loci; P < 2.439*10-8). We identified 87 novel loci, 51 of which were replicated. Furthermore, an extensive list of GWAS follow-up analyses was performed to annotate the underlying genes and pathways and to evaluate the genetic architecture of these fatty acid traits. Notably, a transcriptome-wide association study in the UK Biobank European samples revealed an additional twelve novel loci. Pathway enrichment analysis further highlighted the shared genetic basis between PUFAs and multiple psychiatric disorders, including alcohol use disorders and bipolar disorders. This study improves our understanding of the genetic basis of unsaturated fatty acids and will inform the future practice of gene-based dietary interventions.

Keywords: GWAS; Polyunsaturated fatty acids; omega-3 fatty acids; Genetics;

Flash Talk Session II
Monday, August 8, 2022
11:45 AM - 12:15 PM
Rittenhouse

Identification of genetic loci associated with the risk of aneuploidy with maternal-origin using PGT-A sequences

Siqi Sun[1], Aishee Bag[1], Daniel Ariad[2], Mary Haywood[3], Mandy Katz-Jaffe[3], Rajiv McCoy[2], Karen Schindler[1, 4], Jinchuan Xing[1, 4]

[1]Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA.
[2]Department of Biology, Johns Hopkins University, Baltimore, MD, USA.
[3]CCRM Genetics, Lone Tree, CO, USA.
[4]Human Genetic Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ, USA.

Abstract
Aneuploidy, the inheritance of extra or missing chromosomes, frequently arises during human meiosis and is the primary cause of early miscarriage and maternally age related in vitro fertilization (IVF) failure. The exact genetic causes of variability in aneuploid egg production remain unclear despite discovery of several genetic variants that predispose women to a higher incidence of meiotic aneuploidy. Preimplantation genetic testing for aneuploidy (PGT-A) with low-coverage whole-genome sequencing (lc-WGS) is a standard test for selecting IVF embryos with a normal chromosome complement. The wealth of embryo aneuploidy rate data and lc- WGS data from PGT-A can potentially be used to identify novel loci associated with aneuploidy. By combining lc-WGS data from full-sibling embryos, we imputed genotype likelihoods of genetic variants in parental genomes. We then used these imputed data, as well as aneuploidy calls from the embryos to perform a genome-wide association study of aneuploidy rates. We identified two loci, occurring on chromosome 3 and 9, respectively, associated with maternal meiotic aneuploidy risk. Several candidate genes (e.g., ERC2 and CCDC66) encompassed by these loci are known to be involved in chromosome segregation during meiosis. Together, our work improves understanding of the genetic basis of maternal meiotic aneuploidy risk, while introducing a generalizable method that can be leveraged for similar association studies of lc-WGS data.

Keywords: female fertility, aneuploidy, IVF, preimplantation genetic testing for aneuploidy, low-coverage whole-genome sequencing, association study

---

An R Shiny app for systematically integrating genetic and pharmacologic cancer dependency maps

Tapsya Nayak[1], Li-Ju Wang[2], Michael Ning[3], Yufei Huang[2], Yu-Chiao Chiu[2*], Yidong Chen[1*]

[1]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, TX USA
[2]UPMC Hillman Cancer Center, University of Pittsburgh, PA USA
[3]Department of Computer Science, University of Texas at Austin, TX USA

Abstract
The rapidly growing cancer dependency maps pave the way to precision oncology by identifying and targeting the "Achilles' heel" of cancer. There is a pressing need for software that systematically links such genetic (gene knockouts) and pharmacologic dependencies (small compounds). Here we present an web-based R Shiny app that incorporates heterogenous data from large-scale high-throughput CRISPR screens, pharmacologic screens, and molecular signatures library, jointly covering 17k genes, 20k drugs, and 1k cell lines. The major goal is to match gene knockouts and drug treatments that induce similar effects in cell viability and/or gene expression perturbation in order to address two fundamental questions: 1) which drugs can be potential surrogates to the knockout of a gene, and 2) which genes are potential targets or mechanisms of action of a drug. The app has four complementary and interconnected modules that address various query scenarios to identify potential druggable genetic vulnerabilities and understand the mechanisms of action of a known or new drug. The results are represented by interactive figures and networks, as well as annotated data tables. In summary, our Shiny app enables easy and systematic navigation, visualization, and integration of the rapidly evolving genetic and pharmacologic dependency maps of cancer.

Keywords: Cancer Dependency Maps; High-throughput Screening; R Shiny App; Web Server; Data Visualization and Integration

---

dRFEtools: Dynamic recursive feature elimination for omics

Kynon JM Benjamin[1,2], Tarun Katipalli[1], and Apuã CM Paquola[1,2]

[1]Lieber Institute for Brain Development, Baltimore, MD, USA;
[2]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Abstract
Technology advances have generated larger omics datasets with applications for machine learning. Even so, in many datasets, the number of measured features greatly exceeds the number of observations or experimental samples. Dynamic recursive feature elimination (RFE) provides a flexible feature elimination framework to tackle this problem and to gain biological insight by selecting feature sets that are relevant for prediction. Here, we developed dRFEtools that implements dynamic RFE, and show that it reduces computational time with high accuracy compared to RFE. Given a prediction task on a dataset, dRFEtools identifies a minimal, nonredundant, set of features and a functionally redundant set of features leading to higher prediction accuracy compared to RFE. We demonstrate dRFEtools' ability to identify biologically relevant information from genomic data using RNA-Seq and genotype data from the BrainSeq Consortium. dRFEtools provides an interpretable and flexible tool to gain biological insights from omics data using machine learning.

Keywords: feature elimination, dynamic recursive feature elimination, interpretable machine learning, genomics

---

Genomic Data Augmentation Based on Few-shot Generative Domain Adaptation

Chen Song, Emily Thyrum, Xinghua Shi

Department of Computer and Information Science, College of Science and Technology, Temple University, Philadelphia, PA, USA

Abstract
Despite recent advances in generating large-scale genomic sequences, human genomic data still suffer from data imbalances and biases due to various factors including disease rareness and test affordability. For example, genomic data available are centered around populations with European ancestry and data from other populations are scarce. Hence, this study aims at increasing the quality and volume of genomic data from underrepresented groups by transferring a Generative Adversarial Network (GAN) pretrained on genomes in a majority population to a minority population with extremely small sample size. The main challenge in doing so is to synthesize highly-realistic and diverse data under limited supervision. In this regard, we deploy a few-shot transfer learning strategy to adopt a pretrained model trained on a majority population into another minority population. In particular, we train a GAN stacked with a sequential of onvolutional layers to capture the underlying pattern of the genome from the majority population. We then reuse the prior knowledge from genomics of the majority population by freezing the well-learned low-level layers. Two adaptive layers are employed to learn the high-level diversity in the generator and discriminator respectively. Furthermore, a truncation trick is implemented to constrain the generative space, which is hard to optimize with few training samples. We experimentally evaluate the proposed approach on the cancer genome Atlas (TCGA) prostate gland cancer genotype data from one majority population and two minority populations. Our approach brings appealing results in various settings, substantially surpassing state-of-the-art alternatives, especially in terms of improving the diversity of synthesized data.

Prediction of Pathological Stages in Prostate Cancer Using Graph Attention Networks

Wenkang Zhang[1], Chen Song[1], Zhengkang Fan[1], Xinghua Shi[1*]

[1]College of Science and Technology, Temple University, PA USA

Abstract
Graphs are powerful representations of various types of data including genomic sequences and gene expression. Recently, graph neural networks have gained great attention in machine learning due to its capability of capturing the graphical relationships among data entities. In this study, we develop an efficient graph neural network algorithm to learn the graphical representation of gene expression, with

an aim of predicting pathological stages of prostate cancer using gene expression profiles in patients. We first construct a graph using pairwise dot product quantification values of gene expressions between any two genes to capture the relationships of these genes regarding their similarity in gene expression. Utilizing such a graph, we then leverage graph attention networks to build a predictive model to predict pathological stages of patients. Finally, our graph attention model provides a way to identify marker genes that drive the classification of pathological stages of prostate cancer. Using The Cancer Genome Atlas (TCGA) data, our experimental results show that our proposed method reaches the-state-of-art predicting performance while being effective in identifying marker genes.
Keywords: graph attention networks; cancer classification; prostate cancer

---

Prediction of return of spontaneous circulation using physiological waveforms during cardiopulmonary resuscitation in a pediatric experimental model of cardiac arrest: a machine learning pilot study

Luiz E. V. Silva[1], Lingyun Shi[1], Tiffany S. Ko[2], Hunter Gaudio[2], Vivek Padmanabahn[2], Ryan W. Morgan[2], Julia M. Slovis[2], Rodrigo M. Forti[3], Sarah Morton[2], Yuxi Lin[2], Gerard H. Laurent[3], Jake Breimann[3], Bo Yun[3], Nicolina R. Ranieri[3], Madison Bowe[3], Wesley B. Baker[3], Todd J. Kilbaugh[2], Fuchiang Rich Tsui[1]

[1]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia;
[2]Department of Anesthesiology and Critical Care Medicine, Children's Hospital of Philadelphia;
[3]Division of Neurology, Children's Hospital do Philadelphia.

Abstract
Physiological monitoring during pediatric cardiopulmonary resuscitation (CPR), including invasive hemodynamic variables and end-tidal CO2 (EtCO2), are emerging biophysical metrics of the return of spontaneous circulation (ROSC), the clinical indicator of short-term CPR success. To date, research on hemodynamic guidance of CPR has largely been based on thresholds for single parameters (e.g., systolic blood pressure). The use of multimodal data and the extraction of more complex features in physiological waveforms (i.e., beyond their mean levels) may improve ROSC prediction. In this study, we used machine learning to combine features extracted from eight physiological waveforms to predict ROSC. The waveforms acquired during the first 2 to 10 minutes of CPR in an experimental piglet model of cardiac arrest were analyzed (N = 89 piglets). The waveforms were divided into different time intervals for feature extraction, and logistic regression models were trained on these features for the prediction of ROSC. The receiver operating characteristic curve (AUC) for the combined multivariate model comprising waveform features from the first 10 minutes of CPR was 0.93 [0.87, 0.98] (AUC [95% confidence interval]). This AUC was higher than the AUCs for single waveform models. Both the combined multivariate model and single hemodynamic waveform models provided good predictive performance only when the second half of the CPR period was considered (i.e., after minute 6). We found that waveform features beyond the mean values carried important information for ROSC prediction and could help improve CPR guidance and, subsequently, improve the mortality rate and neurological outcomes after CPR in both research and clinical practice.
Keywords: Return of spontaneous circulation, cardiac arrest, pediatric, piglets, machine learning, waveforms.

Flash Talk Session III
Monday, August 8, 2022
4:40 - 5:10 PM
Grand Ballroom

PRIME Evolutionary Imputation (PREI)

Hannah Kim[12], Sergei L Kosakovsky Pond[12]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA
[2]Department of Biology, Temple University, Philadelphia, PA, USA

Abstract
Introduction: Most nucleotide substitution models treat amino-acid altering, non-synonymous changes as equally likely. In reality, radical amino-acid substitutions often indicate functional changes, and amino-acid physicochemical properties act as constraints to substitutions. In our previous method, PRoperty Informed Models of Evolution (PRIME), we have incorporated amino acid physicochemical properties into a codon substitution model (MG94xREV) to estimate the degree of property importance for each site in a multiple sequence alignment.
Research significance: PRIME Evolutionary Imputation (PREI) is an extension of PRIME. PRIME uses evolutionary history and physicochemical properties to define the sample space in which a site can evolve, and determines whether the property shaping the current space is changing or conserved. PREI expands PRIME by using its estimates to impute credibility for future codons in per-site, per-sample resolution. The new predictive measure can forecast shortrange evolution for a single species from the long evolutionary history across species. It also enables the early detection of functionally important sites under adaptive or conservative selection.
Methodology: A phylogenetic tree can be conditioned on the PRIME maximum likelihood parameters. PREI estimates evolutionary credibility for a sequence by reconstructing marginal probabilities of observing codon i over unobserved ancestral states at the sequence of interest on the phylogenetic tree. In other words, we derive credibility for the sequence of interest with the distribution of possible states from the rest of the sequences on the tree.
Results: On the simulated datasets, PREI shows high overall prediction accuracy. We observed that sequences on the shorter branches tended to show higher accuracy than those on the longer branches.
Conclusion: PREI provides evolutionary credibility for the future states of codons using a model fitted by considering both evolutionary history and physicochemical properties.

Keywords: evolutionary inference; physicochemical properties; computational biology

---

Systems Pharmacogenomic Framework Identifies Associations between Gut-microbiota Metabolites and GPCRome in Alzheimer's disease

Yunguang Qiu1, Yuan Hou1, Yadi Zhou1, Jielin Xu1, James B. Leverenz2,3, Andrew A. Pieper4-7,

Jeffrey Cummings8, Feixiong Cheng1,2,9,*

[1]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[2]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA
[3]Lou Ruvo Center for Brain Health, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[4]Harrington Discovery Institute, University Hospitals Cleveland Medical Center, Cleveland, OH 44106, USA
[5]Department of Psychiatry, Case Western Reserve University, Cleveland, OH 44106, USA
[6]Geriatric Psychiatry, GRECC, Louis Stokes Cleveland VA Medical Center; Cleveland, OH 44106, USA
[7]Department of Neuroscience, Case Western Reserve University, School of Medicine, Cleveland, OH 44106, USA
[8]Chambers-Grundy Center for Transformative Neuroscience, Department of Brain Health, School of Integrated Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA
[9]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

Abstract
Alzheimer's disease (AD) is a progressive neurodegenerative disorder caused by multiple pathophysiological factors, including both genetic and environmental factors. Accumulating evidence suggested that gut-microbiota metabolites play a vital role in mediating AD pathophysiology, despite their potential targets remain unclear. In this study, we developed a systems pharmacogenomics framework that integrates machine learning (ML), AlphaFold2- derived structural biology, and multi-omics approaches to identify disease-relevant metabolites derived from gut-microbiota with non-olfactory G-protein-coupled receptors (GPCRome). Specifically, we evaluated over 1.68 million metabolite-protein pairs connecting 408 human GPCRs and 515 gut metabolites using an Extra Trees algorithm-improved structural pharmacogenomics strategy. Using Mendelian randomization analysis of large AD genetic data, we identified 7 likely causal GPCR targets (e.g., TAS2R60, GALR1 and FPR1) for AD. Using multi-omics (transcriptomics and proteomics) analysis, we identified 10 potential AD-associated GPCRs (e.g., C3AR1). Using three-dimensional structural fingerprint analysis of metabolite- GPCR complexome, we identified over 60% allosteric pockets of orphan GPCR models for gut metabolites in GPCRome, including AD-related orphan GPCRs (e.g., GPR27, GPR34, and GPR84) that were identified as potential targets deorphanized by gut metabolites in AD. Furthermore, we identified the potential targets of two AD-related metabolites (3-hydroxybutyric acid and Indole-3-pyruvic acid) and four mechanism-of-action relevant metabolites from ADrelated bacterium Eubacterium rectale. We further revealed that Tridecylic acid as a disease relevant metabolite for orphan GPR84 in AD. In summary, this study presents a powerful systems pharmacogenomics approach to identify therapeutic potentials from orphan GPCRome, gut-microbiota and multi-omics findings for AD.

Keywords: Alzheimer's disease, Gut Microbiota, GPCRome, systems pharmacogenomics, machine learning, target identification

Tensor-Based Multi-Modality Multi-Target Regression for Alzheimer's Disease Diagnosis

Jun Yu[1], Yong Chen[2], Li Shen[2], Lifang He[1]

[1]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA;
[2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

Abstract

Background: Multi-Target Regression (MTR) has recently attracted great interest in the research community including healthcare, computational medicine, machine learning, etc. The inherent property of jointly learning multiple tasks can contribute task-correlated information to these regressions, thus leading to a better performance than Single-Target Regression (STR). Existing methods focus more on the identification of a specific feature set for different regression tasks, which predefines a suboptimal condition that different tasks are separately modeled but share the common feature space. Thus, how to simultaneously build inter-target correlations and input-target relationships into a task-integrated learning framework is greatly concerned in MTR. In addition, heterogeneous features obtained in a real-world dataset containing multiple modalities or views might complicate this concern via the curse of dimensionality.

Methods: We propose a general Tensor-based Multi-modality MTR (TMMTR) method to address this problem via boosted sparse and low-rank learning. Specifically, we leverage the tensor structure to exploit high-level and inter-target correlation information inherent in the multi-modality multi-target data and investigate tensor-level sparsity in the multilinear regression model. The intrinsic tensor structure is explored to strengthen and capture the complex relationships among multiple regression tasks. In this study, we apply the TMMTR to analyze multimodal imaging data (VBM-MRI, FDG-PET, and AV45-PET) from ADNI cohort with three clinical parameters of Disease Severity (DS) score, AD Assessment Scale–Cognitive 13-item (ADAS-Cog-13) score, and Mini-Mental State Examination (MMSE) score as targets for Alzheimer's disease diagnosis.

Results: The experimental results demonstrate the outstanding performance of our proposed method against both the state-of-the-art STR and MTR methods for the AD diagnosis. The results also show that the significant correlation between the three clinical scores boosts the proposed model's capacity for disease prediction. At the same time, our proposed model learns a better biomarker identification of disease-specific brain regions and modality-related differences from its inherent property of feature selection and competes with other baseline methods in terms of running time. The interpretable coefficients of our model coincide with the clinical findings in AD diagnosis and thus further validate the superiority.

Conclusions: The tensor-structured sparsity, inter-target correlation and input-target relationship are simultaneously exploited to learn the interpretable coefficients in our proposed model, which also successfully identifies biomarkers related to AD with multiple clinical scores and achieves higher predictive performance than the state-of-the-art methods. Our approach is of wide general interest as it can be generalized to other diseases when high dimensionality data is available.

Keywords: Alzheimer's disease, multi-target regression, tensor, feature selection, factorization, interpretability

---

Identifying Chronic Tic Disorder subtypes using clinical diagnostic data

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

[1]Rutgers, the State University of New Jersey, Department of Genetics and the Human Genetics Institute 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 the 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 and Random Forest Classifier, we analyzed the TIC Genetics diagnostic data (36 variables) for 844 CTD probands. Both methods identified the same five distinct clusters: 1. All probands with CTD but did not meet the diagnosis criteria of TS. These probands also have a high prevalence of non-white ancestry; 2. Probands with TS, with low rate of Obsessive-Compulsive Disorder (OCD) and a high prevalence of non-white ancestry; 3. Probands with TS and high rates of OCD and white ancestry; 4. Probands with TS from multi-birth, likely due to environmental conditions during/after birth; and 5. Probands with TS and Trichotillomania and high prevalence of attention deficit hyperactivity disorder (ADHD). In summary, our results show that distinct clusters can be identified among CTD patients based on diagnosis 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

---

Sequences of Events from the Electronic Medical Record and the Onset of Infection

Caitlin E. Coombes, Kevin R. Coombes, and Naleef Fareed

Abstract
Introduction: Dynamic prediction of hospital-acquired infection (HAI) in the intensive care unit (ICU) has important implications to reduce morbidity and mortality. Here, we present a novel model of time-series analysis to learn from electronic health record (EHR) data not only whether an outcome of interest occurred but also when it was likely to have occurred, so as to serve as a basis for future dynamic

intervention. We apply methods translated from the analysis of protein sequences and drawn from Bayesian statistics to predict time-of-onset of serious infection in critically ill patients from sequences of physician actions.
Methods: Using data from Medical Information Mart for Intensive Care (MIMIC)-III for hospitalizations of patients who spent time in an intensive care unit (ICU), we describe each hospital course as a categorical vector, or "alphabet" of 23 events in temporal order from the EMR that we expected to be relevant to the onset of infection during a hospitalization in an intensive care unit (ICU). Translating validated techniques from proteomics, we analyzed these as k-mers of length 3-12 events, condensing very high frequency events identified as network hubs in a graph. For each k-mer that was found in at least 200 patient event sequences, we computed the relative risk (RR) associated with its presence or absence. Finally, we applied a Bayesian model of (cumulative) RR to patient event sequences, updated with information weighted averaging against an uninformative prior to predict time-of-onset of HAI.
Results: For 48,536 hospitalizations with at least 12 events, we computed RR of being prescribed antibiotics (indicating provider identification of HAI) for each k-mer seen in at least 100 cases. The log2-transformed RR (median = 0.248, mean = 0.226) supported the conclusion that the events selected were individually associated with increased risk of infection. Each patient's hospitalization was modeled as an RR curve. Selecting from all possible cutoffs of maximum gain (MG), MG > 0.0244 predicts administration of antibiotics with PPV 82.0%, NPV 44.4%, and AUC 0.706. All positive lead times (i.e. patients for whom antibiotics were administered before the MG point) occurred in patients with only one antibiotics event, suggesting possible prophylactic or pre-procedural use.
Conclusions: Time-series methods are important targets of methods development to improve early identification and prediction of critical clinical syndromes in the ICU. Our approach of translating a well-validated sequence analysis method from bioinformatics to a clinical timeseries problem has not been previously employed. Although this model poses challenges to real-time estimation of risk in the ICU due to lag in updating k-mer length, in its current iteration it holds value for retrospective analysis of other clinical syndromes for which time-of-onset is critical to analysis but poorly marked in the EHR, such as sepsis, delirium, and decompensation.

Sequences of Events from the Electronic Medical Record and the Onset of Infection A pan-cancer analysis and identification of glucose-6-phosphate dehydrogenase (G6PD) inhibitors by cheminformatics approach

Madhu Sudhana Saddala1 and Jiang Qian[1]

[1]Wilmer Bioinformatics Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

Abstract
Glucose-6-phosphate dehydrogenase (G6PD) is commonly considered as the first and rate limiting enzyme of the pentose phosphate pathway (PPP). The pentose phosphate pathway (PPP) has recently been shown to have a crucial role in cancer cell growth by providing both nucleotide precursors, needed for proliferation, and NADPH used for both intracellular ROS (Reactive Oxygen Species) detoxification and catabolic metabolism. G6PD up-regulation can serve as a surrogate marker for cancer staging and is an indicator of a poor prognosis. Nevertheless, the underlying mechanism remains unclear and no pan-cancer analysis is available. Therefore, we first explored the potential oncogenic

roles of G6PD across thirty-three tumors based on the datasets of TCGA (The cancer genome atlas) and GEO (Gene expression omnibus). A comprehensive analysis of G6PD gene expression in tumors, survival prognosis, tumor immunity and immunosuppressive cell infiltration, DNA methylation, gene alteration analysis, and immunotherapy response was performed. We investigated the role of G6PD in the development and prognosis of various cancers. It expression was found to be higher in cancer tissues than in normal tissues in most TCGA cancer types and subtypes and was related to tumor stage, metastasis, and prognosis. Our results linked G6PD to immune cell invasion and immune evasion. The methylation level of G6PD inversely correlated with mRNA expression. G6PD expression levels are associated with immunological and chemotherapeutic outcomes in various cancers. In addition to that we found novel G6PD inhibitors which have different scaffolds, structure-based pharmacophore model was built and validated by different methods. Then, the pharmacophore features were used for chemical databases (ZINC database) for the virtual screening. The selected 27 hit molecules were further evaluated by molecular docking, protein-ligand interactions and in silico ADMET studies. Finally, five compounds with different scaffolds were selected by docking studies with best binding energies. These small molecules were predicted to have high inhibitory activity and good ADMET properties. Therefore, these novel compounds are probably to become a good lead molecule for the development of effective drugs against resistant form of cancers. In the meantime, our pan-cancer study also proposes a relatively comprehensive understanding of the oncogenic roles of G6PD across different tumors.

Keywords: G6PD, pan-cancer analysis, cancer, pharmacophore modeling, cheminformatics.

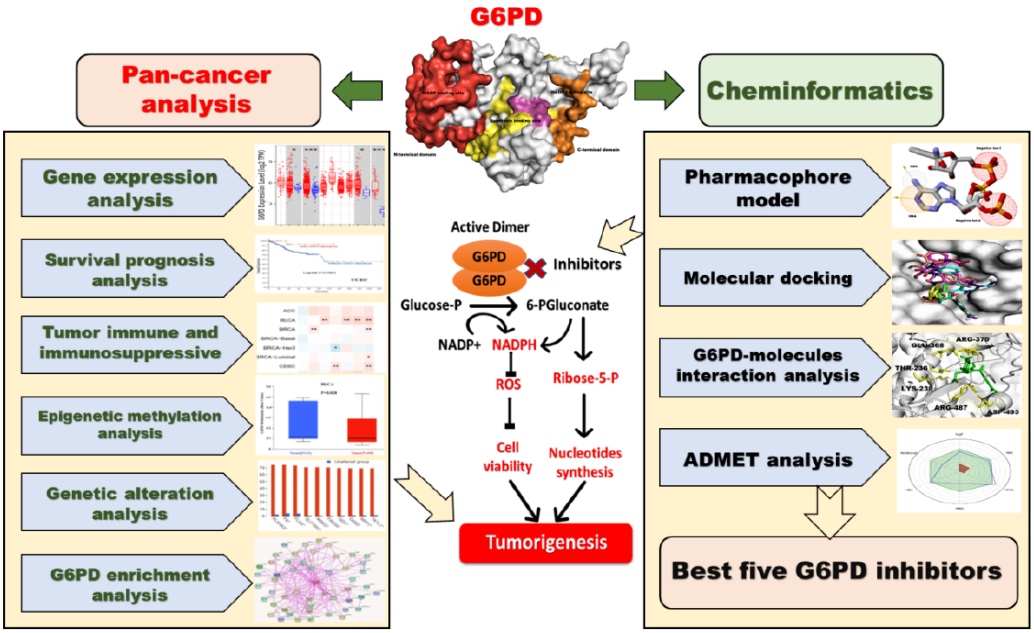


Figure1: The detailed workflow of the present study. Pan-cancer analysis shows gene expression analysis, survival prognosis analysis, tumor immune and immunosuppressive cell infiltration, epigenetic methylation analysis, genetic alteration analysis and functional enrichment analysis. The cheminformatics pipeline indicates pharmacophore modeling, virtual screening, molecular docking and in silico ADMET analysis.

Flash Talk Session IV
Monday, August 8, 2022
4:40 - 5:10 PM
Logan

Identifying drug hepatotoxicity mechanisms using high-throughput concentration dependent toxicity data and toxicokinetic modeling

Daniel P. Russo[1,2], Lauren M. Aleksunes[3], and Hao Zhu[1,2]

[1]Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey, USA
[2]Department of Chemistry, Rutgers University, Camden, New Jersey, USA
[3]Department of Pharmacology and Toxicology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, New Jersey, USA

Abstract
Hepatotoxicity is the injury imposed on the liver caused by exogenous chemicals and is the leading reason of attrition during drug development. Traditional screening for chemical hepatotoxicity relies on animal testing, which is costly, time-consuming, and has ethical concerns. These problems have created a demand for developing higher-throughput alternatives, such as in vitro or in silico models. Models that predict drug adverse outcome pathways (AOPs), the biological events leading to the observed hepatotoxicity, can provide mechanistic insight when evaluating hepatotoxicity. For example, AOP models that leverage responses across multiple related assay targets can simulate and predict the underlying in vivo toxicological processes. In the past two decades, high-throughput screening (HTS) programs have tested millions of chemicals using in vitro assays, providing enormous amounts of publicly available information on toxicity mechanisms. However, using these HTS in vitro assays to develop AOP models is challenging. For example, most assays test biological targets and processes unrelated to hepatotoxicity, and as a result methods are needed to identify and integrate relevant assays into AOP models. Additionally, HTS assays test chemicals at concentrations that may not reflect the in vivo concentration of a drug due to metabolic transformation, making direct prediction of human hepatotoxicity using HTS assays difficult. Here, we present a new computational approach to solve these challenges and establish robust AOP models from concentration dependent HTS assays. First, from a database of 2,171 chemicals with human hepatotoxicity classifications, we obtained concentration response information for 1,609 chemicals from over 1,600 HTS in vitro assays and identified 157 of these to be correlated to human hepatotoxicity. Using a computational framework, these 157 assays were grouped by biological targets (e.g., estrogen receptors) or similar biological mechanisms (e.g., DNA repair) into 52 AOP models of hepatotoxicity. The output of an AOP model is an AOP score summarizing the potency of a chemical against a hepatotoxicity-relevant biological target or mechanism and can be used to rank chemicals. Using these models, the target compounds were ranked and grouped by chemical structure similarity revealing chemical classes with high AOP scores and explained the mechanisms of their hepatotoxicity. Additionally, toxicokinetic models capable of estimating the in vivo human concentrations of chemicals were combined with chemical AOP scores to directly predict in vivo human hepatotoxicity (Accuracy 67%, Recall 64%, Precision

67%). This new computational approach to automatically mine public toxicity data can be a universal strategy for chemical toxicity evaluations.

Keywords: big data, machine learning, computational toxicology, animal alternatives, hepatotoxicity

---

Benchmark study of similarity measures from query phenotypic abnormalities to diseases based on the human phenotype ontology

Yu Hu[1], Kai Wang[1,2]

[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA;
[2]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract
Development of sequencing technologies make it possible for genotype data to be used in clinical diagnosis. However, it is still challenging for clinicians to understand the results of sequencing and make correct judgement based on them. Recent years, phenotype-based diagnosis has been improved with the establishment of the Human Phenotype Ontology (HPO) and the enrichment of phenotype-disease annotations. Here we simulated HPO terms for a group of patients based on 30 complex diseases. We evaluated the performance of disease prediction based on 5 different similarity measures from the HPO terms to hereditary diseases and showed that they consistently achieved high accuracy (>95%) in top 3 candidate diseases. Resnik measure ranked the underlying disease in top 3 on 98.16% of the simulated dataset without noise and 96.32% of the simulated dataset with noise. Second best Jiang-Conrath measure ranked the underlying disease in top 3 on 97.37% of the simulated dataset without noise and 95.85% of the simulated dataset with noise. We also found that all 5 similarity measures provide accurate patient clustering based on simulation study. Our results not only demonstrate the feasibility of phenotype-based diagnosis using existing similarity measures from the HPO terms to hereditary diseases but also highlight necessary bioinformatics improvements for future EHR-based patient clustering tool development in clinical setting.

Keywords: EHR data; HPO term; Similarity measure; Phenotype-based disease diagnosis

---

Functional Impact of Copy Number Variants in Autism Spectrum Disorder and Related Disorders

Rohan Alibutud[1], Vaidhyanathan Mahaganapathy[1], Xiaolong Cao[1,5], Marco Azaro[1], Christine Gwin[1], Sherri Wilson[1], Steven Buyske[2], Christopher W. Bartlett[3], Judy F. Flax[1], Linda M. Brzustowicz[1,4], Jinchuan Xing[1,4,*]

[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; Department of Pediatrics, College of Medicine, The Ohio State University, Columbus, Ohio, USA
[4]The Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[5]Current address: Division of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, China

Abstract
Autism spectrum disorder (ASD) is a neurodevelopmental disorder with a complex polygenic genetic architecture. Large structural variants (SVs), and in particular rare copy number variants (CNVs), have been previously linked to idiopathic ASD. This project is a part of the New Jersey Language and Autism Genetics Study (NJLAGS), a collaborative effort to increase thepowerof ASD gene identification by studying related language impairment (LI) and reading impairment (RI) disorders in multiplex families. The NJLAGS cohort generated genotyping microarray data from which we identified CNVs using multiple calling algorithms. From 522 individuals across 115 families, we filtered and prioritized a set of 14 de novo CNVs (ranging from ~15 kbp to ~5.2 Mbp in size) that are most likely to be contributing to idiopathic ASD in probands. In order to assemble a ranking of the best candidate variants, we employed the program StrVCTVRE, a random forest classifier that predicts pathogenicity of SVs based on an array of traits. We combined StrVCTVRE score with tissue expression, prior annotation, segregation pattern, and overrepresentation analysis to accumulate evidence for ASD causation. The results implicate genes in the mTOR and DAG1 signaling pathways, such as RPTOR, TSC1, POMT1, and LHX3 as potential ASD contributors and as targets for future research.

Keywords: autism, genomics, neurodevelopmental disorder, family cohort, copy number variants

---

Immuno-therapy-induced gene signatures in clear cell renal cell carcinoma

Ye-Lin Son[1], Huihui Fan[1], Zhongming Zhao[1,2,3]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]MD Anderson Cancer Center University of Texas Health Graduate School of Biomedical Sciences, Houston, TX 77030, USA

Abstract
Among different types of immune-therapy, immune checkpoint blockade (ICB) represents one of the most promising treatment options that have revolutionized the treatment of various types of human cancers, such as clear cell renal cell carcinoma (ccRCC). Despite the fact that the response of ICB treatment has been largely studied within the tumor microenvironment, studies on understanding its effects on tumor cell trajectories are still lacking.

Here, we applied machine-learning-based trajectory inference on tumor cells subject to ICB treatment, that were profiled using single-cell RNA sequencing from eight ccRCC patients. Three lineages were formed among 7,468 tumor cells, with one of the lineages clearly representing responsive cellular phenotype, while another lineage indicating a more aggressive non-responsive phenotype of tumor subpopulation. By analyzing genes dynamically altered along both lineages, we identified four signature gene sets with distinct expression patterns over trajectory pseudotime points. In line with their lineage-specific responsive or non-responsive phenotypes, we showed that these four gene sets were significantly associated with the disease specific prognosis patterns using The Cancer Genome Atlas ccRCC cohort (n=533) based on single-sample gene set enrichment analysis. In addition, we also conducted a drug-target enrichment analysis using our signature gene sets. Cabozantinib and Epicatechin gallate were among the top-ranked drugs that targeted gene sets dynamically associated with the non responsive tumor cell lineage. These drugs thus highlighted the potential of combinatory use with the ICB treatment in ccRCC patients for enhanced outcome. Altogether, our machine learning based lineage analyses on tumor cells are much needed for a better understanding of tumor heterogeneity in response to ICB treatment, which also paves the way for potential combinatory drug uses in clinical settings.

Keywords: single-cell RNA sequencing; drug response; immune checkpoint blockade; trajectory analysis; prognosis

---

Synchronized decoding of functional capacities and compositions of metagenomes in a sweep

Daniel Roush[1,2], Daniel Hakim[3,4,5], Antonio González[3,4,5], George Armstrong[3,4,5], Justin Shaffer[3,4,5], Daniel McDonald[3,4,5], Rob Knight[3,4,5 *], Qiyun Zhu[1,2 *]

[1]Center for Fundamental and Applied Microbiomics, Biodesign Institute, Arizona State University, Tempe AZ, USA;
[2]School of Life Sciences, Arizona State University, Tempe AZ, USA;
[3]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA;
[4]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, CA, USA;
[5]Department of Computer Science & Engineering, University of California San Diego, La Jolla, CA, USA

Abstract
Microbiome studies enable the identification of functional capacities of communities and their connections to the host health. This is achieved through mapping high-throughput sequencing data against a reference gene database with pre-annotated functions. Functional analysis is independent from interrogating community composition, which instead maps reads against reference genomes. The diversification of the two analyses leads to extra computational expense and engineering challenges, as well as limited capability of connecting functional to host microbial genomes, which is essential for understanding microbial roles in biological processes.
We propose the combined analysis of microbiome function and composition using one single alignment of sequencing reads against reference genomes. This integration relies on the use of genome coordinates. Inspired by the sweep line algorithm, we developed an efficient algorithm that performs

a single iteration over an ordered sequence of start and end positions of both genes and reads to identify overlap. It achieves an exact solution with close to linear time complexity. It is scalable to modern large datasets with millions of reads, thousands of genes and databases of hundreds of reference genomes. We implemented this algorithm in the software package Woltka, with features for conducting functional and taxonomic classification and summarization of metagenomic data. Meanwhile, we expanded the reference database WoL to contain a comprehensive catalog of pre-annotated classification units and hierarchies. The method has been adopted in Qiita and has analyzed tens of thousands of samples.
We evaluated the performance of Woltka and compared it to current methods. Using simulated metagenomes and the WoL database, we showed that read-gene associations extracted from genome alignment are more accurate than the classical gene alignment, calculated using both nucleotide and translated gene alignments. We next compared Woltka with HUMAnN 3,the current state-of-the-art metagenomic functional profiling software, and showed that Woltka outperforms the latter in both computational efficiency and assignment accuracy, using different reference databases (WoL and the ChocoPhlAn database) and measured on different functional annotation levels (UniRef, KO, and EC). Analysis of real-world datasets such as the inflammatory bowel disease cohort of the Integrative Human Microbiome Project (iHMP) showed that Woltka reveals biological signals as well or moderately better than HUMAnN 3, with higher consistency between taxonomy and function, at a much lower computation cost.
We demonstrated that Woltka effectively utilizes genome alignment to achieve robust functional classification of metagenomes, and therefore recommend Woltka as part of the data analysis workflow for microbiome studies.

Keywords: Metagenomics, functional annotation, sweep line algorithm, microbiome

---

A novel water aware QM-MM hybrid method for macromolecule drug discovery and bio-active conformational prediction

Yuxin Xie[1], Wanting Chen[1], Shengpei Chen[1]

[1]Shenzhen xNA Biotechnology Co., Ltd. Shenzhen, Guangdong, CHINA.

Abstract
Conformational prediction of biological macromolecule reveals its ability to target or being targeted, and bring powerful insights to drug discovery. Many computational methods have been developed to serve this purpose, such as, quantum mechanics (QM)-based small molecule structural property with ab initio calculation, molecular mechanics (MM)-based protein secondary structure folding with data driven deep neural network, and molecular dynamic based solvent-solute interaction with all-atom simulation algorithm.
However, the crucial solvation effect is ignored by most QM applications to tolerate the compute-savvy process; and the lacking of theoretical accuracy in MM leads to unreliable prediction. All of these defects slow down current drug design and discovery, especially therapeutic macromolecules.
In this study, we first established a novel water aware QM-MM hybrid method, which specialized in bio-active macromolecular structural prediction for drug discovery. Different from typical QM and

MM, our method includes a comprehensive hybrid geometry optimization cycles. During which, a short water-aware QM learning step is used for the system parameterization; and then the updated parameters will be fed into a long learning MM step to simulate a solvent interaction for next cycle's water-awareness. A divide-conquer algorithm is recruited to further reduce our computing cost.
Additionally, we discovered discrete aqueous orbits around PO/PS of oligonucleotides backbone, which is highly associated with oligonucleotides' charge distribution, polarity and cell affinity. Also, we observed a significant structural diversity of PO/PS bond, which indicates our first principle method could truly reflect the molecular configuration and be beneficial to therapeutic macromolecule discovery.

Keywords: quantum mechanics, molecular mechanics, drug discovery, therapeutic macromolecules, solvation effect, conformational prediction.

Poster Session

Sunday, August 7, 2022
5:45-7:30 PM

Abstract ID: 34
Transfer learning to predict functional impact of missense variants by language model embedding of protein sequences

Alan Tian[1], Yige Zhao[2], Yufeng Shen[2,3]

[1]Lynbrook High School, San Jose, CA, USA
[2]Department of Systems Biology, Columbia University, New York, NY, USA
[3]Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract
The uncertain effect of missense variants complicates the identification of risk genes in both clinical diagnosis and genetic studies. Accurate prediction of the functional impact of a missense variant is critical in genome sequence interpretation. Although many computational models were available to predict pathogenicity of missense mutations, their performance varies. Recent advances in new machine learning models of protein sequences and large population genome data sets present new opportunities to improve the accuracy of computational prediction. Here we developed a pathogenicity predictor to evaluate the utility of protein language models in predicting functional impact of missense variants. We obtained a pre-trained protein sequence model (ESM-1b) and a multi-sequence alignment model (ESM-MSA). For any missense variant of interest, we computed embedding vectors of flanking protein sequence through ESM-1b and ESM-MSA, and concatenate these as the input to a 5 layer feed-forward classification model. We obtained training data from curated pathogenic variants databases, such as ClinVar and Human Gene Mutation Database, in total 59,701 positives and 59,701 negative missense variants. From these, 20,000 variants were randomly selected to train a classifier with an 80/20 train test split. With the test data, the model was able to achieve an AUC of 0.82 using raw amino acids, and 0.77 using MSA embeddings. When both embedding vectors are appended together and trained on a classifier, it achieves an AUC of 0.80. These results are comparable with other models, such as PrimateAI (0.79), MPC (0.81), and CADD (0.80). In the future, an end-to-end model focused specifically on mutational effects may be used to further improve the performance.

Abstract ID: 50
MetaRNN: Differentiating Rare Pathogenic and Rare Benign Missense SNVs and InDels Using Deep Learning

Chang Li, Degui Zhi, Kai Wang and Xiaoming Liu

Abstract

With advances in high-throughput DNA sequencing, numerous genetic variants have been discovered in the human genome. One challenge we face is interpreting these variants to help in disease screening, diagnosis, and treatment. While multiple computational approaches have been proposed to improve our understanding of genetic variants, their ability to identify rare pathogenic variants from rare benign ones is still lacking. Using context annotations and deep learning methods, we present pathogenicity prediction models, MetaRNN and MetaRNN-indel, to help identify and prioritize rare nonsynonymous single nucleotide variants (nsSNVs) and non-frameshift insertion/deletions (nfINDELs). A recurrent neural network incorporating a window of +/- 1 codon around the affected codon was combined with 28 high-level annotation scores and allele frequency features to develop the two proposed models. Employing independent test datasets, we demonstrate that these new models outperform state-of-the-art competitors and achieve a more interpretable score distribution. Importantly, prediction scores from the nsSNV-based and the nfINDEL-based models are comparable, enabling easy adoption of integrated genotype-phenotype association analysis methods. In addition, we provide pre-computed MetaRNN scores for all possible human nsSNVs and a Linux executable file for a fast one-stop annotation of nsSNVs and nfINDELs. All the resources are available at http://www.liulab.science/MetaRNN.

Abstract ID: 52

DeepCore: Attention-based interpretable deep learning approach for detecting regulatory elements

Hai Chen[1,2], Pramod Chandrashekar[1,2], Navid Ahmadinejad[1,2], Li Liu[1,2]

[1]College of Health Solutions, Arizona State University, Phoenix, AZ, 85004, USA
[2]Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, AZ, 85281, USA

Abstract

In this study, we report a novel computational method, DeepCore, to discover various types of REs and elicit their tissue specificities.

Unlike existing algorithms that are trained to classify REs vs. non-REs without considering interdependence of genetic elements, DeepCore identifies multiple REs of a gene concurrently by assessing their joint impact on gene transcription.

It first builds an attention-based deep neural network to predict transcription level of a gene in a single sample based on epigenetic profiles of 100kbps regions surrounding the transcription start site, then combines neural attentions derived from multiple samples to identify REs.

Using experimentally validated data as benchmarks, we show that DeepCore can accurately predict enhancers and silencers with AUROC values of 0.81 and 0.87, respectively, which outperforms existing methods. We show that the DeepCore is general framework to integrate multi-omics data of heterogeneous samples to derive biologically interpretable machine learnin models.

Keywords: Regulatory elements, Deep learning, Attention, Histone modification, Multiomics

---

Abstract ID: 53

Predicting sequence specificities of experimentally unexplored RNA-binding proteins via optimizing the evolutionary correlation

Shu Yang[1*], Jiahang Sha[1*], Kefei Liu[1], Sumita Garai[1], Jingxuan Bao[1], Zixuan Wen[1], Raymond T. Ng[2], Li Shen[1]
*These authors contributed equally to this work.

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

Abstract
In order to understand the complex and abundant interactions between RNAs and proteins in the cell, an essential step is to characterize the binding specificities of RNA-binding proteins (RBPs) to the target RNAs. Many RBPs are reported to display intrinsic preferences to specific nucleotide sequence motifs despite that RNAs can naturally fold into various structures. Although wet-lab experiments have generated a large amount of data to study the RNA-RBP interactions by far, only a small subset of RBPs have their binding specificities been experimentally explored. The binding information on the vast majority of the RBPs is still unknown.
In this work, we propose to tackle the challenges by utilizing the correlated evolutionary relationship between RBPs and their target RNA sequence motifs. Intuitively, similar in spirit to the setting of zero-shot learning, here we leveraged the existing binding data from experimentally explored RBPs to infer the binding sequence specificities for RBPs without binding data, using evolution as auxiliary information. Specifically, we set the evolutionary correlation as the objective function and formulated the prediction of sequence specificities as optimization problems. We took the position weight matrix (PWM) as our sequence specificity representation and evaluated our method on the benchmarking CISBP-RNA database. The preliminary results showed that our predictions are better than the previous nearest neighbor alternatives and comparable to the true specificity model.
In summary, while experimentally characterizing the binding specificities of RBPs is challenging and expensive, our method may serve as a simple workaround to provide insights on the bindings of unexplored RBPs. For the next steps, we plan to extend and apply our method to study the cross-talks between RNAs and proteins in some important cellular processes like disease-related post-transcriptional regulations and virus infections.

Keywords: RNA-binding proteins, Evolution, Optimization, Data mining

---

Abstract ID: 54

Recent Deep Learning Studies for Microfluidic Assays of Yeast Lifespan

Hong Qin[1], Mehran Ghafari[2], Weiwei Dang[3]

[1]Department of Computer Science and Engineering, University of Tennessee at Chattanooga, Chattanooga, TN, USA

[2]Del E. Webb Center for Neuroscience, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, California
[3]Huffington Center on Aging, Baylor College of Medicine, Houston, TX, USA

Abstract
We wish to highlight our deep learning methods to facilitate the microfluidic-based high throughput assay of the replicative life spans of the budding yeast [1-3]. The budding yeast is a proven model of cellular aging and has revealed many longevity-related genes with conserved roles in eukaryotic cells. Microfluidics-based assays have become an effective way to monitor the aging process for a large number of cells, resulting in hundreds of time-lapse microscopic images. We are the first groups to demonstrate that the deep learning method can accurately classify cell division events [1]. We developed a maximal likelihood approach to infer the cell family tree and cell lineages [2]. We showed that YOLO and Mask-RCNN are complementary for cell object detections [3]. We will compare our results with two pre-prints, including one from a CALICO/GOOGLE research group. Overall, it is promising that deep learning methods could potentially transform cellular aging research.

Keywords: deep learning, microfluidics, budding yeast, replicative lifespan

---

Abstract ID: 58
Discoveray of DNA methylation protector element in human genome

Jingmin Shu[1,2], Jaroslav Jelinek[1], Yan Zhan[1], Hai Chen[2], Jean-Pierre J. Issa[1], Li Liu[2]

[1]Department of Leukemia, UT. M. D. Anderson Cancer Center
[2]Department of Biomedical Informatics, Arizona State University

Abstract
Aberrant DNA methylation is a hallmark of cancer, and promoter hypermethylation has been associated with silencing of tumor suppressor genes (TSG) and plays a key role in tumorigenesis.Previous work suggests that spreading of heterochromatin from methylation centers may account for the hypermethylation around TSG promoters, although the genetics basis of methylation spreading and protection is unclear. To address this problem, we curated a large collection of genomic regions that contains repetitive DNA elements, CpG islands and gene promoters, hereafter referred as Repeat-Island-TSS (RIT) Trio elements. By integrative analysis of multi-omics data from normal and cancerous colon samples, we have identified more than one thousand RITs with DNA methylation pattern consistent with spreading from methylation center into promoters in cancer but protected in normal tissues and was able to identify a highly conserved DNA element that is a candidate of protective motif against methylation spreading. In ex vivo experiments we have been able to illustrate that this motif indeed protects against methylation spreading, while mutations to this motif lead to loss of the protection. Above results for the first-time shed light on the mechanisms of dysregulation of DNA methylation in cancer and may lead to further clinical applications in diagnosis and treatment of cancer.

Keywords: DNA methylation, CpG islands, Multi-omics

Abstract ID: 60
Large-scale surface protein imputation in single cells with transfer learning

Ruoqiao Chen[1], Jiayu Zhou[2], Bin Chen[1,2,3]

1Department of Pharmacology and Toxicology, Michigan State University, Michigan, USA;
[2]Department of Computer Science and Engineering, Michigan State University, Michigan, USA.
[3]Department of Pediatrics and Human Development, Michigan State University, Michigan, USA.

Correspondence: Bin Chen. E-mail: chenbi12@msu.edu

Abstract
Surface proteins are crucial for studying cell functions and are increasingly explored as diagnostic or therapeutic targets. Experimental quantification of thousands of membrane proteins in single cells proves to be difficult. CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing), the most advanced technology to quantify surface proteins' abundance, is very expensive and only capable of measuring at most 300 surface proteins at present. The mRNA expression of surface proteins derived from single cell (sc)RNA-Seq is thus often used as a surrogate; however, the scarcity of CITE-seq data and the inconsistency between mRNA and protein expression levels call for a better means of protein abundance estimation.
Since transcripts encode protein information and are much more accessible, we develop a novel computational framework that predicts protein abundance based on the expression of existing transcripts and their prior relations. The framework includes data normalization, feature embedding and sophisticated deep-learning architecture. Transfer learning is leveraged to enable prediction across various tissues. CITE-seq data are used to refine and verify models.
By utilizing available CITE-seq data for only ~200 proteins, our model can predict the abundance of >1000 cell surface proteins from any given human single-cell transcriptomes, which significantly expands the scale of surface protein abundance data. With transfer learning, our model successfully made prediction in blood, bone marrow, brain, lung, and pancreas. For the prediction on the proteins seen in the training set, our model achieved comparable performance with existing models such as cTPnet and Seurat V3. More importantly, with an improved average correlation between prediction and ground truth (cor: 0.34) compared to the correlation between RNA raw counts and surface protein abundance (cor: 0.26) in an external validation, our model is the first to be capable of predicting proteins absent from the training set. Visualization of the imputed protein abundance values by our model defines clearer cell opulations. In patients, with imputed protein abundance our model identifies cell-specific surface markers that are overlooked by the approaches using only scRNA-Seq data (e.g., CD279 in T cells in glioma).
In conclusion, we propose a promising approach for imputing large-scale surface protein abundance from single-cell transcriptome data sets. Such predicted protein abundance values could further facilitate the study of specific cellular functions in human diseases and the identification of new immunotherapy targets by providing clearer cell population classification and more accurate quantification of cell-specific surface markers.

Keywords: Transfer learning; Single cell RNA-seq; CITE-seq; Surface protein quantification; Cancer immunotherapy

---

Abstract ID: 61

Giant chromosome, enhancer co-amplification, and altered 3D genome structure contribute to oncogenic gene expression in liposarcoma

Tingting Liu[1*], Juan Wang[1*], Hongbo Yang[1], Qiushi Jin[1], Xiaotao Wang[1], Yihao Fu[1], Yu Luan[1], Qixuan Wang[1], Mark W. Youngblood[2], Xinyan Lu[3], Lucia Casadei[4], Raphael Pollock[4,5#], Feng Yue[1,6#]

[1]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine Northwestern University, Chicago, IL, USA.
[2]Department of Neurosurgery, Feinberg School of Medicine Northwestern University, Chicago, IL, USA.
[3]Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.
[4]Program in Translational Therapeutics, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
[5]Department of Surgery, The Ohio State University, Columbus, OH 43210, USA
[6]Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, IL, USA.

[*]These authors contributed equally.
[#]Corresponding authors.

Abstract

Liposarcoma (LPS) is the most common soft-tissue sarcoma in adults with two major subtypes, well differentiated and dedifferentiated. Both subtypes are characterized with the pathognomonic giant ring or marker chromosomes that harbor high copy-number of known oncogenes. Here, we reported a comprehensive molecular characterization of both tumor and normal tissues from the same LPS patients, including WGS, transcriptome, enhancer landscape, and genome-wide 3D genome structure by Hi-C. We identified tumor-specific transcripts and regulatory elements, and discovered enhancer co-amplification and hijacking events as novel mechanisms upregulating the oncogenes such as MDM2, CDK4 and HMGA2. By combining Hi-C, optical mapping, nanopore long reads and WGS data, we partially resolved complex structure variations (SVs) and reconstructed the local genome and the giant chromosome. Overall, our study provides the comprehensive resource for LPS and offers novel insights of how altered enhancers and 3D genome contribute to gene dysregulation in cancer.

Keywords: Liposarcoma, enhancer hijacking, enhancer co-amplification

---

Abstract ID: 66

Identifying stage-specific imaging genetic patterns in Alzheimer's disease

Daniele Pala[1], Brian Lee[1], Xia Ning[2], Dokyoon Kim[1], and Li Shen[1]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Biomedical Informatics, College of Medicine, The Ohio State University

Abstract
Alzheimer's disease (AD) is one of the most common and severe forms of Senile Dementia. Genome-wide association studies (GWAS) have identified dozens of AD susceptible loci. To better understand a potential mechanism-of-action for AD, quantitative brain imaging features have been studied as mediators linking genetic variants to AD outcomes.
Mediation analysis and mixed-effects models are used to investigate the biological pathways by which genetic variants affect both brain structures/functions and disease diagnosis. We analyzed the imaging and genetics data collected from the ADNI project, including a Polygenic Hazard Score (PHS) as an indication of genetic risk, and 13 imaging quantitative traits (QTs) extracted from the AV45 PET scans quantifying the amyloid deposition in different brain regions. The participants were from four diagnostic groups: Normal (NL), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD.
Mediation analysis assessed the mediating effects of image QTs between PHS and diagnosis. The relation between each PHS and each mediator was assessed using a standard linear regression, whereas a logistic regression was used for measuring the relation between the PHS and diagnoses, with different diagnostic comparisons (i.e., NL vs. non-NL, NL+EMCI vs LMCI+AD, non-AD vs. AD) conducted. The results show that all 13 imaging QTs have a significant mediating effect in separating NL and non-NL subjects, a majority of mediators (9 out of 13) have a significant effect splitting NL+EMCI and LMCI+AD participants, and only two mediators are significant in detecting the difference between non-AD and AD.
Linear mixed-effects models were used to characterize intra-group differences in the associations between genetic scores and imaging QTs for different disease stages. Random intercepts were calculated for each of the four diagnostic groups and for all imaging QTs separately and were all significant, with higher values for more advanced disease stages.
In summary, our mediation analysis and mixed-effects models have identified promising stage-specific imaging QTs that mediate the genetic effect of the studied PHS on disease status. These results provide novel insights into the predictive power of the PHS and the mediating power of amyloid imaging QTs with respect to multiple stages over the AD progression.

Keywords: Alzheimer's disease, imaging genetics, mediation analysis, mixed-effect model

---

Abstract ID: 68
Explorations in Specifying Loss Functions and Meta-Structures for DNA Methylation via Deep Predictive Modelling

Houyi Du, Zeinab Parsons, Jin Lu[1]

[1]Department of Computer and Information Science, College of Engineering and Computer Science, University of Michigan, Dearborn, MI, USA

Abstract
Machine Learning (ML) has been successfully applied to DNA methylation prediction. Several studies show that deep learning has great performance compared with other methods. To achieve good generalization performance across a wide variety of cell types, the choice of the loss function and the meta structure of the deep learning-based systems matter, because it implicitly assumes a particular error distribution and feature extraction structure. However, because the implications of the choice of loss function are not obvious, the selection is often challenging and not discussed thoroughly. In our study, via various single cell-type and multi-cell-type experiments, we pinpoint the characteristics of the loss functions that are linked to the generalization errors, therefore, guiding the loss function and meta structure selection process in different cases when data is either abundant or relatively limited. Through investigating regions like CpG island, we discovered that the hinge shows better effectiveness especially when the sample size is limited, or the samples are from a single cell type since it focuses training on a sparse set of hard examples and prevents most easy examples from overwhelming the prediction model.
When data is abundant and from multiple cell types, to mitigate the negative transferring problem during training, we also propose a novel meta-structure for multi-task learning called Dynamic Weight Average with Moving Average (DWAMA), which enables the model to attend to hard-to-predict tasks/cell types. Our results show that when trained with the DWAMA, the prediction accuracy could increase the attention to hard tasks hence surpassing the accuracy of the other multi-task learning meta-structures.

Keywords: DNA methylation, Multitask Learning, Loss function

Abstract ID: 69

Exploring Automated Machine Learning for Alzheimer's Disease Classification and Amyloid Imaging Biomarker Discovery using STREAMLINE

Boning Tong[1], Yanbo Feng[1], Xinkai Wang[1], Marylyn Ritchie[1], Jason Moore[2], Ryan Urbanowicz[2], Li Shen[1]

[1]University of Pennsylvania, Philadelphia, PA, USA
[2]Cedars-Sinai Medical Center, West Hollywood, CA, USA

Abstract
Motivation: Alzheimer's disease (AD) is an irreversible neurodegenerative disease, and a timely and accurate AD diagnosis can lead to improved intervention and treatment. Many machine learning models have been studied for AD classification. However, it is not straightforward to perform intuitive performance comparison among different models. STREAMLINE (https://github.com/UrbsLab/STREAMLINE) is an automated machine learning (AutoML) pipeline that can quickly analyze data using different modeling algorithms while adhering to ML best practices. In this work, we use STREAMLINE to explore and compare the prediction / feature selection performances of 12 binary classification models for AD diagnosis and biomarker discovery.

Method: Participants (N=377) included 213 cognitively normal (CN) and 164 AD subjects from the ADNI cohort. To classify CN vs AD diagnostic status, we employed 116 regional amyloid imaging measures extracted from AV45 PET scans. Our STREAMLINE analysis included four steps: (1) preprocessing and feature transformation, (2) feature importance evaluation and selection, (3) modeling, and (4) postprocessing. We evaluated twelve commonly used classification algorithms, including Naive Bayes, Logistic Regression, Random Forest, SVM, Artificial Neural Network, etc. Training was conducted using three-fold cross-validation (CV) with a Bayesian sweep optimizing model hyperparameters for each algorithm/training set combination. Model predictive performance was compared by evaluating balanced accuracy, ROC AUC, Precision-Recall curve (PRC) AUC, running time, and the feature importance for all 12 algorithms. The pipeline generated a comprehensive report summarizing and comparing the above performance measures of 12 methods and the corresponding feature importance maps through various figures and tables.
Results and conclusion: Among 12 classification models, the logistic regression algorithm yielded the highest balanced accuracy (0.850), ROC AUC (0.915), and PRC AUC (0.908) with less running time, which made it the best model for AD classification with AV45 PET scans. After consolidating all the amyloid imaging feature importance results, hippocampus and precuneus appear among the top regions contributing to prediction. This was consistent with the results of MultiSURF, an algorithm that conducts feature importance evaluations independent of modeling. In summary, these results demonstrate the feasibility and effectiveness of STREAMLINE as an AutoML pipeline for evaluating AD diagnostic classification models, and for discovering AD amyloid imaging biomarkers.

Keywords: Alzheimer's disease, AutoML, amyloid imaging, classification, biomarker discovery

---

Abstract ID: 71

Genome-Wide Association Study using Novel Tiling Array Identifies Genomic Subsequences Associated with Alzheimer's Disease Quantitative Traits

Brian Lee[1], Jingxuan Bao[1], Sarah Wait Zaranek[2], Jiong Chen[1], Yuhan Cui[1], Junhao Wen[1], Shu Yang[1], Heng Huang[3], Andrew J. Saykin[4], Paul M. Thompson[5], Dokyoon Kim[1], Christos Davatzikos[1], Alexander Wait Zaranek[2], Li Shen[1], for the ADNI and AI4AD initiatives

[1]University of Pennsylvania, Philadelphia, PA, USA
[2]Curii Corporation, Somerville, MA, USA
[3]University of Pittsburgh, Pittsburgh, PA, USA
[4]Indiana University, Indianapolis, IN, USA
[5]University of Southern California, Marina del Rey, CA, USA

Abstract
Traditionally, genome-wide association studies (GWAS) have analyzed SNP array or whole genome sequencing (WGS) data. Although useful in identifying individual variants associated with a quantitative trait or disease phenotype, the linear nature of these representations makes analysis of sets of variants difficult. Therefore, to facilitate the joint analysis of multiple combinations of nearby SNPs, a novel genomic tiling array (https://curii.co/su92l-j7d0g-swtofxa2rct8495) was proposed for genetic studies using statistical and machine learning methods.

In this work, we used genomic tiling, imaging and clinical data from 1,337 individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Before modeling, quality-control checks were performed for each tile (a consistently named subsequence in the genome) and its respective variants. Ordinary least-squares (OLS) linear regression models were used to evaluate the association between a tile variant and a quantitative trait of interest (AV45, or CDR-SB score). Clinically relevant covariates also factored in the models included sex, years of education, age, and ancestry (as represented by the first 10 ancestry principal components). Association strength was quantified via a P value; statistical significance was determined via a Bonferroni correction.
Regression models identified four tile variants significantly associated with both quantitative traits. The four tile variants were from the APOE and APOC1 regions, where genes APOE and APOC1 have been linked to AD pathology. These findings confirmed the viability of a tiling-based approach to GWAS in localizing AD-related genetic regions. These promising associations motivate future investigations to explore the potential of the genomic tiling representation in machine learning studies.

Keywords: Genome-wide association study, genomics, tiling array, Alzheimer's disease

---

Abstract ID: 80

Multi-dimentional Precise Drug Screener (MPDS): a webserver for precise in sillico drug screening

Jiannan Liu[1], Tianhan Dong[2], Huanmei Wu[3], Kun Huang[4], Jie Zhang[5]

[1]Dept of BioHealth Informatics, Indiana University School of Informatics and Computing, Indianapolis, IN, USA.
[2]Dept of Pharmacology and Toxicology, Indiana University School of Medicine, Indianapolis, IN ,USA.
[3]Dept of Health Services Administration & Policy, Temple University College of Public Health, Philadelphia, PA, USA.
[4]Dept of Biostatistics & Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA.
[5]Dept of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA.

Abstract

As new drug design and development have been a resource intense and costly process with low success rates, the fast and low cost in sillico drug repurposing has gained increasing popularity. The L1000-based Connectivity Map (CMAP), one of the largest dataset available today that provides systematic evaluation of transcriptomic change induced by small moleculars, can be used as a rich resource for such drug screening and repurposing. In this study, we designed and implemented Multi-dimentional Precise Drug Screener (MPDS), a webserver that can be used to perform customizable drug screening using features directly or inferred from transcriptomic data of CMAP.
MPDS integrates cancer cell line related information from CCLE database and gene expression profiles from CMAP before/after drug perturbation, it provides rich measurements of cancer cell lines and CMAP instances, such as cancer cell line's genome alterations inferred from CCLE copy number data,

stemness levels of CMAP instances measured by PanCanStem and pathway activities of CMAP instances measured by ssGSEA. The user interface of MPDS is systematically designed and implemented to provided a guided and intuitive user experience. For biomedical researchers, MPDS can be used to profile drug candidates for certain disease with specific conditions such as lung cancer with chromosome 3q amplification. The diverse measurements of CMAP instances enables drug profiling with various creitera, for example, users can profile drugs that decresase the WNT pathway activities in lung cancer and can also profile drugs that decrease stemness level in breast cancer. The development of MPDS eased biomedical researchers' effort of performing in sillico drug screening using publicly available datasets, it will accelerate the drug repurposing research and can potentially lead to the discovery of precision treatment using available drugs to cure currently untreatable disease.

Keywords: Drug repurposing, drug screening, MPDS, CMAP, CCLE

Abstract ID: 83
Augmenting Hi-C data using Generative Adversarial Networks (GANs)

Chong Li, Chen Song, Xinghua Shi

Department of Computer & Information Sciences, Temple University, Philadelphia, PA, USA

Abstract
Techniques like high-throughput chromosome conformation capture sequencing (Hi-C) have been widely used to characterize the three-dimensional (3D) structure of the genome and uncover folding principles of chromatin. Chromatin conformational structures such as topologically associating domains (TADs) provide information to investigate contacts of regulatory elements and genes that are subsequently linked to human diseases. TADs are reported to play a strong role in insulating genes from aberrant regulation introduced by regulatory elements outside TADs (such as enhancer hijacking), while TAD boundaries can prevent the spreading of transcription and repressive chromatin. Given that high-resolution Hi-C data is hard to ascertain and is thus less abundant comparing with other types of sequencing data, we leverage recent advances in data augmentation using Generative Adversarial Networks (GANs) to significantly improve the volume and quality of Hi-C data. Specifically, we develop a GAN framework for Hi-C data augmentation utilizing a combined stack of convolutional layers in both generator and discriminator of a GAN model. Experimental results on published Hi-C data demonstrate that our model is capable of accurately generating higher quantity of high-resolution Hi-C matrices.

Keywords: Hi-C, GAN, chromatin 3D structure, TADs, TAD boundaries

---

Abstract ID: 86
Exploring the Peripheral Kynurenine Metabolites as an Alzheimer's Disease Biomarker: A Systematic Review and Meta-analysis

Mehmet Enes Inam, BS[1], Brisa S. Fernandes, MD, PhD[1], Zhongming Zhao, MSc, PhD[1,2,3]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2]Faillace Department of Psychiatry and Behavioral Sciences, McGovern Medical School, The University of Texas Health Science Center at Houston (UTHealth), Houston, TX 77030, USA
[3]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Abstract
Introduction: The kynurenine pathway has been increasingly attracting attention as a relevant pathway in neurological disorders, including Alzheimer's disease (AD). Early detection of such metabolic alterations may allow for timely intervention and management, potentially decreasing treatment failure rates. Here, we conducted a systematic review and meta-analysis of the kynurenine pathway metabolites from blood samples in AD.
Methods: PubMed and EMBASE databases were searched from journal inception to April 2022 to identify peer-reviewed case-control studies that assessed kynurenine metabolites in AD compared to healthy controls (HC) in peripheral blood, namely, tryptophan (TRP), kynurenine (KYN), kynurenic acid (KA), quinolinic acid (QA), and 3-hydroxykynurenine (3-HK). The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed. The random effects model parameter was selected when comparing the standardized mean differences (SMD) between groups.
Results: Data were extracted from 17 articles that met the inclusion criteria. The mean sample size for each study for individuals with AD and HC was 32±25 and 37±36, and the mean age was 74.1±10.5 and 72.7±9.2, respectively. TRP levels (n=16) were decreased in AD (SMD: -0.83, CI95% [-1.18 to -0.49], $p<0.001$). KA levels (n=8) were also decreased (SMD: -0.27, CI95% [-0.46 to -0.08], p=0.005). There were no significant differences in KYN (n=10), 3-HK (n=7), and QA (n=5) levels, but a significant increase in the KYN/TRP ratio (n=7) was observed in AD compared to HC (SMD: 0.55, CI95%[0.04 to 1.05], p=0.030).
Conclusion: Tryptophan, the precursor of the kynurenine pathway, is decreased in individuals with AD. KA, which has neuroprotective effects, is also decreased in AD. The KYN/TRP ratio is increased, suggesting a shift in tryptophan degradation toward the kynurenine pathway. Drugs that target the kynurenine pathway, particularly KA, might be useful in AD treatment. Future studies should validate and explore the utility of this peripheral biomarker as a treatment target.

Keywords: Alzheimer's disease, Biomarker, Metabolomics, Molecular pathophysiology, Cellular metabolism, Neuroscience

---

Abstract ID: 90
Quantifying modality-specific brain signatures for cognitive traits in Alzheimer's disease

Zixuan Wen[1*], Jingxuan Bao[1*], Shu Yang[1], Mansu Kim[2], Andrew J. Saykin[3], Yize Zhao[4], Li Shen[1], and for the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Artificial intelligence, Catholic University of Korea, Bucheon, Republic of Korea
[3]Indiana Alzheimer Disease Center, Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA
[4]Department of Biostatistics, School of Public Health, Yale University, New Haven, CT, USA

*These authors contributed equally to this work

Abstract

Introduction: Inspired by the definition of heritability, morphometricity is a recently introduced concept to measure the variability of any given trait that can be attributed to morphometric variation. It has been studied to quantify MRI-based brain morphometric signature for cognitive traits. In this work, we generalize this concept to "brain-X-metricity", which measures the variability of any given trait that can be attributed to brain variation captured by a specific neuroimaging modality X. We perform a proof-of-concept study to estimate "brain-X-metricity" of multiple cognitive traits with respect to three imaging modalities (i.e., X=MRI, AV45, or FDG) using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) biobank.

Methods: Participants included over 900 non-Hispanic Caucasian subjects with complete baseline measurements of studied imaging and cognitive trait information. We focused on analyzing three imaging modalities: structural MRI (measuring brain morphometry), AV45-PET (measuring amyloid burden), and FDG-PET (measuring glucose metabolism). For each modality, 116 regional measures were calculated by averaging all the voxel-level measures within each brain region defined by the AAL atlas. We examined five cognitive traits: Alzheimer's Disease Assessment Scale-Cognitive (ADAS13), Clinical Dementia Rating (CDRSB), Functional Activities Questionnaire (FAQ), Mini-Mental State Exam (MMSE), and Rey Auditory Verbal Learning Test (RAVLT-learning). Brain-X-metricity estimation is implemented using Linear Mixed Effects (LME) model, where the imaging-based subject similarity matrix is calculated using Gaussian metric.

Results: Our results indicated that the variability of all five traits were highly or moderately attributed to all three modality-specific brain signatures. MMSE yielded the highest brain Xmetricity estimates (i.e., 1 for AV45 and FDG, and 0.96 for MRI). RAVLT.learning yielded the lowest estimates (i.e., 0.5 for AV45, 0.49 for FDG, and 0.42 for MRI). The other three traits yielded intermediate and similar estimates (0.79-0.83 for AV45, 0.7-0.84 for FDG, 0.57-0.76 for MRI).

Discussion and conclusion: Our brain-X-metricity analysis demonstrates that MMSE has the strongest brain signatures captured by all three modalities. Each of five cognitive traits has stronger brain-AV45-metricity and brain-FDG-metricity than brain-MRI-metricity. These brain- X-metricity findings capture valuable high-level imaging-cognition associations, which can be used to prioritize and guide subsequence investigations on identifying detailed associations between regional brain measures and cognitive traits.

Keywords: Brain-X-metricity; modality-specific brain signature; multimodal imaging; cognitive trait; Alzheimer's disease

---

Abstract ID: 91

AlphaFold 2 Monomer: deployment in an HPC environment

Yuntao Yang, Zhao Li, David J. H. Shih, W. Jim Zheng

School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.

Abstract
AlphaFold2, developed by Google DeepMind, is a breakthrough in the grand challenge of protein structure prediction. While the breakthrough will have profound impact on biomedical research, its application faces significant hurdles due to the computing intensive nature. We overcome this challenge by deploying the AlphaFold 2 pipeline in an HPC environment that fully utilized the computing resources and accelerated the workflow. Specifically, the CPU component of the AlphaFold 2 that includes multiple sequence alignment and template search was deployed on a computer cluster at the Texas Advanced Computing Center (TACC). The
high performance of CPU cores and I/O requests on the cluster allowed us to complete over 200 jobs within 10 hours. The GPU component that includes model prediction and refinement was deployed on the latest Nvidia GPU server, and 100 jobs could be completed within 24 hours when 2 jobs run in parallel. The deployed workflow can efficiently use different computing environments to process many protein structure predictions to advance biomedical research.

Keywords: AlphaFold 2, HPC, Nvidia GPU

Abstract ID: 96
Individualized connectomic predictive biomarkers of antidepressant and placebo responses in major depression

Kanhao Zhao[1], Xiaoyu Tong[1], Hua Xie[2], Nancy Carlisle[3], Yu Zhang[1]

[1]Department of Bioengineering, Lehigh University, Bethlehem, PA, USA
[2]Department of Psychology, University of Maryland, College Park, MD, USA
[3]Department of Psychology, Lehigh University, Bethlehem, PA, USA

Abstract
Sertraline, commonly prescribed in major depressive disorder patients (MDDs), exhibited marginal superiority over placebo. This is partly caused by the individuals' neurobiological heterogeneity. To better dissect the heterogeneity, and then define more precisely personalized treatment-predictive signatures, the individual-unique functional architecture of the brain might be helpful. We established informative signatures of treatment responses to antidepressant medication and placebo in MDD using individualized functional connectivity (FC). The data were collected from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care study. Subjects were blindly and randomly assigned to the placebo or sertraline treatment for 8 weeks. The whole-brain FC networks were constructed from pre-treatment resting-state functional magnetic resonance imaging (fMRI). We then applied the common orthogonal basis extraction technique to extract individualized FC of each

subject. LASSO was trained with the individualized FC features to predict treatment response measured as pre- minus 8-th week change in the total score of 17-item Hamilton Depression Rating Scale (HAMD17). The predictive models successfully identified signatures that explained 22% variance for the sertraline group and 31% variance for the placebo group in predicting HAMD17 change. Compared with the raw FC-based models, the individualized FC-defined signatures significantly improved the prediction performance, as confirmed by the 10 x10-fold cross-validation (Wilcoxon signed-rank test result of R2 difference; sertraline: w = 2.57, pfdr = 0.01; placebo: w = 3.02, pfdr = 0.004). For the sertraline treatment, the predictive FCs were predominantly located in the left middle temporal cortex and right insula. For the placebo treatment, the predictive FCs were primarily involved in the bilateral cingulate cortex and left superior temporal cortex. The right prefrontal lobe was the shared important region for both treatment response predictions. After FC individualization, the predictive weights from FCs in those regions, and the Pearson correlation coefficient between these FCs with the treatment response were enhanced. Our results suggested that the individualized FCs successfully enlarged individual-unique behavior-related brain dysfunction, which defined novel treatment-predictive signatures in MDD. Additionally, the important regions involved in predictive FCs for sertraline and placebo treatment were generally different, however, lots of these regions played a key role in mood regulation, thereby the placebo and antidepressant effects might target similar brain functions with different brain connections.

Keywords: Individualized functional connectivity, Antidepressant, MDD, fMRI

Abstract ID: 97

Association and Risk Factor Identification between Periodontitis and Alzheimer's Disease using Electronic Dental Record Data

Jay Patel[1], Bari Dzomba[1], Xinghua Shi[2], Domenico Pratico[3], Huanmei Wu[1]

[1]Department of Health Services Administrations and Policy, College of Public Health, Temple University, Philadelphia, PA, USA
[2]Department of Computer & Information Sciences, College of Science and Technology, Temple University, Philadelphia, PA, USA
[3]Alzheimer's Center at Temple, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA

Abstract

Introduction: Alzheimer's disease (AD) is a chronic neurodegenerative disease with dementia and a critical health problem associated with the aged population, reaching almost 50% incidence for those aged 85 and older. Periodontitis is a widespread gum disease due to gum and bone infections and inflammation with bacteria. Preliminary studies have demonstrated there could be potential associations between AD and periodontitis. For instance, host inflammation is a common factor between AD and periodontitis. However, there is no systematic study to understand the underlying associations and the common risk factors linking the two diseases.

Data and Methods: This study looks at two cohorts of patients with periodontal disease, i.e., the periodontitis with AD (Perio+AD) group and periodontitis without AD (Perio-AD) group, using the

dental electronic health record (EHR) of patients at Temple University Kornberg School of Dentistry between January 1, 2017, and August 31, 2021. Natural language processing is performed on clinical notes to identify patients who reported AD during their dental visits. Various potential risk factors, such as patient demographic characteristics, teeth conditions, lifestyle behaviors, and other dental conditions, were retrieved from the EHR using an automated program. The periodontitis stages were generated using the American Academy of Periodontology classification systems with Stages I-IV and Grades A-C. The associations between AD and periodontitis are compared between the two patient cohorts using statistical analysis, including t-test and chi-square tests.
Results: Preliminary comparison of 141 'Perio+AD' patients and randomly selected 141 'Perio-AD' patients demonstrated that 'Perio+AD' patients had a significantly higher incidence of Stage IV periodontitis (19% versus 6%) and dental pain (14% versus 3%). However, our analysis showed no significant differences in the prevalence of Stage I, stage II, stage III, gingivitis, chewing function, decayed teeth, missing filled teeth, speaking, and dental aesthetic. Demographic information analysis showed that 'Perio+AD' patients were older than Perio-AD (61 years versus 52 years), although there was no significant difference between patients' gender or race in these two cohorts.
Conclusions: Our study demonstrated an association between AD and severe periodontitis (Stage IV) with underlying common risk factors. Further studies are needed on large populations to confirm these findings. These findings will be valuable for policymakers to provide advanced dental care to patients with AD. Moreover, it is essential to investigate whether periodontitis is responsible for AD initiation and progression to take preventive approaches.

Keywords: Alzheimer's Disease, Periodontal Disease, Periodontitis, Electronic Health Records

---

Abstract ID: 103
Normative Modeling of Functional Connectivity for Antidepressant Response Prediction

Xiaoyu Tong, Hua Xie, Nancy Carlisle and Yu Zhang

Abstract
Antidepressant medications yield unsatisfactory treatment outcomes in patients with major depressive disorder (MDD) with modest advantages over placebo. This modest efficacy is partly due to the elusive mechanisms of antidepressant responses and unexplained heterogeneity in patient's response to treatment — the approved antidepressants only benefit a portion of patients, calling for personalized psychiatry based on individual-level prediction of treatment responses. Normative modeling, a framework that quantifies individual deviations in psychopathological dimensions, offers a promising avenue for quantifying individual-level heterogeneity and achieving the personalized treatment for psychiatric disorders. In this study, we built a normative model with resting-state electroencephalography (EEG) connectivity data from healthy controls of three independent cohorts. Specifically, an autoencoder was employed to capture the expected variability within healthy subjects. Afterward, the individual deviations of MDD patients were calculated as the reconstruction errors yielded by the autoencoder, which were hypothesized to represent the variability in psychopathological dimensions. Using these individual deviations as input features, we trained sparse predictive models for treatment responses of MDD patients, where treatment response was quantified as pre- minus post-treatment changes in the 17-item Hamilton Depression Rating Scale (HAMD17). The antidepressant

treatment lasted for 8 weeks and treatment arms included sertraline and placebo. As a result, we successfully predicted treatment outcomes for both patients receiving sertraline ($r = 0.45$, $p < 0.001$) and placebo ($r = 0.33$, $p < 0.001$), which significantly outperformed the predictions derived from standard EEG connectivity (sertraline: Fisher's $z = 1.84$, $p = 0.033$; placebo: Fisher's $z = 2.84$, $p = 0.002$), demonstrating the advantage of individual deviations to personalized response prediction endowed by normative modeling framework. From the individual deviation-based predictive models, we identified key EEG connectivity signatures in resting-state EEG for antidepressant treatment, indicating the importance of precuneus for sertraline response and the importance of alterations in cortical and limbic regions for placebo response. Together, our findings and highly generalizable framework advance the neurobiological understanding in the potential pathways of antidepressant responses, enabling more targeted and effective MDD treatment.

Abstract ID: 107

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

Abhishek Majumdar, Aida Yazdanparast, Huanmei Wu, Lang Li and Lijun Cheng

Abstract

We present a transfer learning co-module matching model, called, CoMatch for prognostic/predictive biomarker identification. 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.

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, co-match modules to doxorubicin consists of 21 genes, 11 of the genes are from mRNA expression ABCB11, SEMA4F, HERC2, C15orf54, CCDC33, FSTL4, GABRA5, POPDC2, BSND, IL3RA and SLC6A2 and 9 from CNV: MTOR, ANGPTL7, FBXO2, AGTRAP, MTHFR, MAD2L2, FBXO6, FBXO44 and PTCHD2 and MYO3A from mutation in this co-module pattern of cancer cells and patients.

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.

Keywords: breast cancer, machine learning, cancer cells, drug screening, precision medicine

---

Abstract ID: 108

Automated optic disc and cup segmentation for glaucoma detection from fundus images using the Detectron2 framework

Fengze Wu[1], Marion Chiariglione[2], Xiaoyi Raymond Gao[1,2,3]

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA;
[2]Department of Ophthalmology and Visual Sciences, The Ohio State University, Columbus, OH, USA;
3Division of Human Genetics, The Ohio State University, Columbus, OH, USA.

Abstract

Glaucoma is a chronic, degenerative optic neuropathy and a leading cause of blindness worldwide. Individuals with glaucoma do not show typical symptoms for years and can become advanced before patients notice an extensive visual field loss. Therefore, early detection and treatment are crucial to prevent vision loss from this blinding disease. Vertical cup-to-disc ratio (VCDR), the ratio of vertical diameter of cup over vertical diameter of disc in the optic nerve head region, is an important structural indicator for glaucoma. Estimation of VCDR requires accurate segmentation of optic disc (OD) and optic cup (OC) on fundus images. However, manual annotation of disc and cup area is time-consuming and is subjective to personal experience and opinion. In this study, we proposed an automated deep learning approach for OD and OC segmentation and VCDR derivation from fundus images using Detectron2, a state-ofthe-art object instance segmentation framework. We trained Mask R-CNN models for OD and OC segmentation and VCDR evaluation. We assessed the performance of our method on the Retinal Fundus Glaucoma Challenge (REFUGE) dataset in terms of the Dice index (DSC) for OD and OC, and the mean absolute error (MAE) for VCDR. Our method achieved highly accurate results with a DSC of 0.9622 for OD, a DSC of 0.8870 for OC, and an MAE of 0.0385 for VCDR on the hold-out testing images. This surpassed all REFUGE teams by improving OD and OC DSC over top performers by 0.2% and 0.4%, respectively, and reducing the VCDR MAE by 7%. Our method provided an accurate and automated solution for OD and OC segmentation and VCDR estimation.

Keywords: Glaucoma, Fundus Image, Object Detection, Instance Segmentation, Detectron2, Deep Learning

---

Abstract ID: 114

The Intersections of Demographics, Social Vulnerability, and HIV: Assess Health Disparities for People Living with HIV

Huanmei Wu[1], Jay Patel[1], Feifan Liu[2], Ben S. Gerber[3], Omar Martinez

[1]Department of Health Services Administrations and Policy, College of Public Health, Temple University, Philadelphia, PA, USA
[2]Department of Population and Quantitative Health Sciences University of Massachusetts Chan Medical School, Worcester, MA, USA
[3]Department of Social Work, College of Science and Technology, Temple University, Philadelphia, PA, USA

Abstract

Introduction: HIV is a disease embedded in social and economic inequity. Achieving the UNAIDS 95-95-95 goals, i.e., increasing the current levels of HIV diagnosis, treatment, and HIV viral suppression to 95% each, requires addressing social vulnerabilities that contribute to poor engagement in HIV care. Social vulnerability refers to the potential negative health effects on communities caused by external stresses. People living with HIV (PLWH) are at greater risk of housing insecurity/homelessness, high medical costs, limited income, and limited ability to continue working due to co-occurring conditions which in turn, decrease the likelihood that PLWH will adhere to their medication regimen and achieve HIV viral load suppression. Therefore, we aim to study the health disparities impacting people living with HIV using large EPIC Cosmos data at TUHealth.

Methods: We used Cosmos (2017-2022) at TUHealth to create two study cohorts using their ICD codes (one includes TUHealth patients vs. all HIV patients). Patient demographics (such as race, ethnicity, language, insurance, and smoking) and CDC's social vulnerability index (SVI) in socioeconomic status, unemployment, and housing/transportation were compared and analyzed. Based on the SVI percentiles, we categorized social vulnerability into four groups (Low Vulnerable: <25%, Medium Low Vulnerable: 25-50%, Medium Highly Vulnerable: 50-75%, Highly Vulnerable: >75%). The retrieved data is analyzed to examine the association between diagnosed HIV infections and demographics or SVI to assess health disparities. The retrieved patients of each CDC SVI group for each demographic characteristic are compared and tested using paired t-test and Chi-Square test to identify the independence and difference between the two cohorts.

Results: We compared patients with HIV infection (N=7,365) with all active patients (N=867,508). The two patient groups have no significant statistical difference ($p>0.05$) for ethnicity and language. However, significant differences ($p<0.05$) for race, medical insurance, smoking, unemployment SVI, socioeconomic SVI, and housing/transportation SVI was found. Black/African Americans had a significantly higher prevalence of HIV infection than Whites. PLWH were more often on Medicaid but less for smoking. Compared to all TUHealth patients,PLWH were more likely to be included in the CDC's SVI (highly vulnerable) in the categories of socioeconomic status, unemployment, and housing/transportation.

Conclusion: The results show significant statistical differences between the two patient groups guiding the development and prioritization of interventions to address social vulnerabilities and gaps in the HIV care continuum among systemically and structurally excluded populations. Further, efforts are needed to better capture other stressors contributing to social vulnerabilities among PLWH

Keywords: HIV, Data quality, Social determinants of health, EPIC, Cosmos, Electronic health records

Abstract ID: 116

Challenges in heterogeneous medical data from multi-EHR Systems: A case study

Huanmei Wu[1], Jay Patel[1], Gabriel Tajeu[1], Ilene Hollin[1], Recai Yucel[2]

[1]Department of Health Services Administrations and Policy, College of Public Health, Temple University, Philadelphia, PA, USA
[2]Department of Epidemiology and Biostatistics, College of Public Health, Temple University, Philadelphia, PA, USA

Abstract

Introduction: Health informatics research has gained tremendous attention on utilizing nationwide multi-system multi-institute data to generate practice-based evidence for decision making and problem-solving. However, immense challenges exist in integrating these vast datasets with interoperability issues.

Data and Approaches: We obtained patient demographic and clinical data from the integrated HealthShare Exchange (HSX) Clinical Data Repository across the Greater Philadelphia/Delaware regions. The datasets are gathered from 377 healthcare organizations in the last six years containing 44 clinical variables for patient demographics (e.g., age, gender, race, ethnicity), encounters (e.g., visit reason, chief complaint, admission type, discharge information, encounter notes), attending physicians, vitals, diagnosis, and procedures. We performed data quality measures and descriptive statistics on limited variables.

Results: The dataset consisted of 200,341 unique patients with >65 million hospital encounters. These patients had 171.2 million unique vital readings, 150.3 million diagnosis information, and 392.9 million performed procedures. Even though HSX has integrated real-time interoperability solutions, we discovered the following challenges to work with the data.

• Data standard issues across multiple EHR systems (e.g., five distinct coding systems (CSs) for vitals, 17 CSs for diagnosis & 20 CSs for procedures).

• Inconsistent labels, such as >100 diagnosis labels for 'abdominal aneurysm' and >800 labels for 'abdominal pain'.

• Spelling errors, such as 'pain' was misspelled as 'paiin', 'paian', 'paing', 'poain', 'pais', 'pait', 'paim', and etc.

• Discrepancy in documentation, such as different units for vitals (e.g., 'cm', 'in', 'ft' for height), and incorrect/inconsistent race/ethnicity documentations (e.g., 'Hispanic' was documented in race, which should be in ethnicity)

• Heterogeneous synonyms, for example, Hispanic (ethnicity) was represented as 'H,' 'HS, 'Hisp ', 'Hispanic', 'Cuban, 'Puerto Rican ', 'Hispanic or Latino ', 'Mexican ', 'Mexican American ', or 'Chicano,' '2135-2,' and others, along with spelling errors.• Missing values, such as missing 13% of vital CSs, 6% of procedure CSs, and 97% of diagnosis CSs.

Some of these challenges can be addressed using informatics methods. Inconsistent documentation and spelling errors can be resolved using natural language processing. Data standard issues can be addressed by conducting in-depth reviews of electronic health record (EHR) systems, especially about the underlying coding principles. Missing data can be addressed by developing advanced informatics phenotyping algorithms and imputation methods.

Conclusions: Before utilizing EHR data for research, data quality measures must be performed. Data quality issues can be reduced using advanced informatics methods; however, errors made during the data collection process may be difficult to solve.

CONFERENCE LOCATION

Notary Hotel

21 N Juniper St, Philadelphia, PA 19107

Listed on the National Register of Historic Places and boasting sophisticated, 1920s-inspired décor and furnishings, Notary hotel is a remarkably well-maintained landmark in the heart of Center City. Business and leisure travelers alike will appreciate our spacious, recently renovated accommodations, all of which feature complimentary high-speed Wi-Fi, luxurious beds, pristine marble bathrooms and expansive workstations. In addition, there are over 100 Asian and Western style restaurants within walking distance of the hotel.

Parking Information

Parking Garage located on site. Additional Parking:

Parkway Corp Garage – 24 hours
1201 Filbert St. (access from Filbert, 13th, and 12th Streets) Self Park Only
Five Star Parking Lot – 24 Hours
1301 Market St. (corner of Market and 13th St., across from the Marriott) Self Park Only (215) 523-5740
Wanamaker Garage – 24 Hours
1300 Market St. Self Park Only

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

Philadelphia Intercontinental Airport (PHL) - located 12 miles from Hotel, 30-60 minute drive time.

Hotel Information

Notary Hotel
Situated in the heart of Philadelphia.

Map and Directions
Room Rate: $ 169 Double or King

SPECIAL ACKNOWLEDGEMENTS

We are grateful for the numerous helps from the following volunteers:

Mehmet Enes Inam	University of Texas Health Science Center at Houston, TX, USA
Jiajinlong (Jack) Kang	University of Texas Health Science Center at Houston, TX, USA
Mohammad Erfan Mowlaei	Temple University, PA, USA
Chong Li	Temple University, PA, USA
Wubin Ding	Children's Hospital of Philadelphia, PA, USA
Jonathan Perdomo	Children's Hospital of Philadelphia, PA, USA
Sam Chen	UTHealth at Houston
Yu Hu	Children's Hospital of Philadelphia, PA, USA
Umair Ahsan	Children's Hospital of Philadelphia, PA, USA
Andy Wang	Peddie School, NJ, USA
Yunyun Zhou	Children's Hospital of Philadelphia, PA, USA
Mani Subramanian	Rutgers, The State University of New Jersey, NJ, USA
Siqi Sun	Rutgers, The State University of New Jersey, NJ, USA
Rohan Alibutud	Rutgers, The State University of New Jersey, NJ, USA

MANY THANKS TO OUR SPONSORS!

10x GENOMICS®

vizgen

ClinChoice
The Standard of Excellence

BGI
GLOBAL GENOMIC SERVICES

About 10x Genomics

10x Genomics is a life science technology company building products to interrogate, understand and master biology to advance human health.

Our mission

We deliver powerful, reliable tools that fuel scientific discoveries and drive exponential progress. Cited in more than 3,300 research papers, our innovative single cell, spatial, and in situ technologies enable discoveries across oncology, immunology, neuroscience, and more. Our talented, dedicated science professionals have a distinguished record of creating innovative instruments, reagents, and software that analyze biological systems at a resolution that matches the complexity of biology.

For more information please visit www.10xgenomics.com

vizgen

Explore New Dimensions Through Spatial Context with MERSCOPE™, the Premier Platform for Spatial Genomics

The MERSCOPE Platform is powered by MERFISH technology, providing exceptional resolution from whole tissue to subcelluar levels for *in situ* single-cell spatial genomics. The platform's high sensitivity detects even lowly expressed genes with a sequencing-free, automated workf ow.

Available now, discover how the MERSCOPE Platform can advance your research.

ClinChoice

The Standard of Excellence

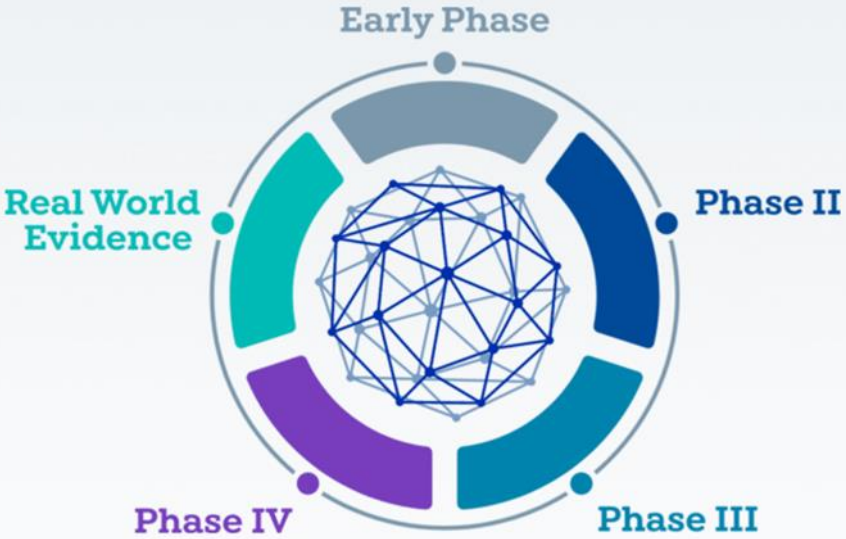
• 25+ Year History of Quality • Flexible Approach • Expedited Timelines • Global Resourcing

Your Full Service Clinical CRO Partner

Pharma | Biotech | Med Device | Cosmetics | Consumer Healthcare

Services for the Full Development Lifecycle

Accelerating drug and device approvals to market with post-market support for more than 25 years



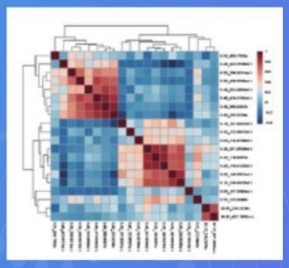
Contact us

Info@ClinChoice.com

215-283-6035

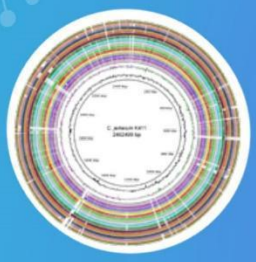
www.clinchoice.com

Formerly FMD K&L



Untargeted Metabolomics

Biomarker discovery
- Well developed workflow
- Large in-house metabolite database



Whole Genome Sequencing

Complete portfolio
- Standard
- PCR free
- Low-pass (lpWGS)
- Long fragment read (lfrWGS)

10X Genomics Single Cell RNA-Seq

Packaged service
- Death cell removal
- 3'RNA lib prep
- Sequencing
- Data analysis

BGI

We offer one of the industry's most comprehensive **multi-omics service portfolios** including proteomics, metabolomics, and much more! All with excellent data quality and quick turnaround time. We transform research and innovate biological discovery.

Label Free DIA

Quantitative proteomics
- High reproducibility
- Customized spectral library
- Label free quantitation

Isobaric Label TMT

Quantitative proteomics
- Accurate and deep
- Global quantitation of proteome

Premium RNA-Seq and Pre-Made Library Sequencing

Coming soon
- Locally delivered
- Rapid TAT
- Accepting pilots