2021 International Conference on Intelligent Biology and Medicine (ICIBM 2021)

August 08-10, 2021
Virtual via Zoom

Hosted by:
The International Association for Intelligent Biology and Medicine (IAIBM),
Temple University,
The Perelman School of Medicine, University of Pennsylvania,
and
The University of Texas Health Science Center at Houston

TABLE OF CONTENTS

Welcome to ICIBM 2021!

On behalf of all our conference committees and organizers, we welcome you to the 2021 International Conference on Intelligent Biology and Medicine (ICIBM 2021), co-hosted by The International Association for Intelligent Biology and Medicine (IAIBM), Temple University, and the Perelman School of Medicine at the University of Pennsylvania. Given the rapid innovations in the fields of bioinformatics, systems biology, and intelligent computing and their importance to scientific research and medical advancements, we are pleased to once again provide a forum that fosters interdisciplinary discussions, educational opportunities, and collaborative efforts among these ever growing and progressing fields.

We are proud to have built on the successes of previous years' conferences to take ICIBM 2021 to the next level. This year, our keynote speakers include Drs. James S. Duncan, Chunhua Weng, Ben Raphael, and Ying Xu. We also have four eminent scholar speakers from Drs. Yue Feng, Graciela Gonzalez-Hernandez, Kai Tan, and Wei Chen. These researchers are world-renowned experts in their respective fields, and we are privileged to host their talks at ICIBM 2021. Throughout the conference, we will feature speakers in four workshops and tutorials that will each provide an in-depth presentation or lesson on some of the most popular informatics topics in the biological and biomedical areas. We have faculty members, postdoctoral fellows, PhD students and trainee level awardees selected from a substantial number of outstanding manuscripts and abstracts that span a diverse array of research subjects. These researchers, chosen through a rigorous review process, will showcase the innovative technologies and approaches that are the hallmark of our featured interdisciplinary fields and their related applications.

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 2021's program. We'd like to extend our thanks to our sponsors for making this event possible, including National Science Foundation, the University of Texas Health Science Center at Houston, and School of Biomedical Informatics. Furthermore, our sincerest thanks to the members of all our committees and our volunteers for their valuable efforts; we could not have accomplished so much without your dedication to making ICIBM 2021 a success.

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 our attendees. Thank you for attending ICIBM 2021. We look forward to your participation in all our conference has to offer!

Sincerely,

Xinghua Mindy Shi, PhD
ICIBM Program co-Chair
Associate Professor
Department of Computer & Information Sciences
College Of Science and Technology
Temple University

Li Shen, PhD, FAIMBE
ICIBM Program co-Chair
Professor of Informatics
Department of Biostatistics, Epidemiology & Informatics
Perelman School of Medicine
University of Pennsylvania

Kai Wang, PhD
ICIBM 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
ICIBM 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

Program Chairs

Xinghua (Mindy) Shi, Temple University, USA
Li Shen, University of Pennsylvania, USA

Steering Committee

Yidong Chen, University of Texas Health Science Center at San Antonio, USA
Xiaohua Tony Hu, Drexel University, USA
Kun Huang, Indiana University, USA
Jason Moore, University of Pennsylvania, USA
Sudhir Kumar, Temple University, USA
Yi Xing, University of Pennsylvania, USA

Program Committee

Xinghua (Mindy) Shi, Co-Chair, Temple University, USA
Li Shen, Co-Chair, University of Pennsylvania, USA
Abolfazl Doostparast, Children's Hospital of Philadelphia, USA
Aimin Li, Xi'an University of Technology, China
Boshu Ru, Merck & Co. Inc., USA
Chaochun Wei, Shanghai Jiao Tong University, China
Chunhua Weng, Columbia University, USA
Cuncong Zhong, University of Kansas, USA
Daifeng Wang, University of Wisconsin, USA
Dajiang Zhu, University of Texas at Arlington, USA
Danielle Mowery, University of Pennsylvania, USA
Dokyoon Kim, University of Pennsylvania, USA
Dong Si, University of Washington, USA
Fuhai Li, Washington University in St. Louis, USA
Garam Lee, University of Pennsylvania, USA
Guimin Qin, Xidian University, China
Haixu Tang, Indiana University, USA
Hanwei Zhang, Northshore University, USA
Honghuang Lin, Boston University School of Medicine, USA
Huanmei Wu, Temple University, USA
Hyunhwan Jeong, UT Health Science Center at Houston, USA
Jiandong Wang, University of South Carolina, USA

Jiang Qian, Johns Hopkins University, USA
Jianhua Ruan, The University of Texas at San Antonio, USA
Jianhua Xuan, Virginia Tech, USA
Jianlin Chen, University of Missouri Columbia, USA
Jiayin Wang, Xi'an Jiaotong University, China
Jim Zheng, School of Biomedical Informatics, U Texas Health Science Houston, USA
Jinchuan Xing, Rutgers, The State University of New Jersey, USA
Jingwen Yan, Indiana University-Purdue University Indianapolis, USA
Jun Wan, Indiana University School of Medicine, USA
Junbai Wang, Radium Hospital, Norway
Junfeng Xia, Anhui University, China
Junjie Chen, Harbin Institute of Technology at Shenzhen, China
Juntao Guo, University of North Carolina at Charlotte, USA
Kefei Liu, University of Pennsylvania, USA
Kun Huang, Indiana University School of Medicine, USA
Kwangsik Nho, Indiana University, USA
Lei Du, Northwestern Polytechnical University, China
Lei Wei, Roswell Park Comprehensive Cancer Center, USA
Lei Xie, City University of New York, USA
Leng Han, Texas A&M University, USA
Li Chen, Indiana University School of Medicine, USA
Lifang He, Lehigh University, USA
Lijun Cheng, Ohio State University, USA
Limin Jiang, Tianjin University, China
Mansu Kim, University of Pennsylvania, USA
Matthew Hayes, Xavier University of Louisiana, USA
Min Xu, Carnegie Mellon University, USA
Min Zhao, University of the Sunshine Coast, Australia
Mirjana Maletic-Savatic, Baylor College of Medicine, USA
Peilin Jia, Beijing Institute of Genomics, China
Ping Zhang, Ohio State University, USA
Qin Ma, Ohio State University, USA
Ranadip Pal, Texas Tech University, USA
Rendong Yang, University of Minnesota Twin Cities, USA
Rich Tsui, Children's Hospital of Philadelphia, USA
Rui Kuang, University of Minnesota Twin Cities, USA
Rui Xiao, University of Pennsylvania, USA
Rui Zhang, University of Minnesota Twin Cities, USA
Ruli Gao, Houston Methodist, USA
Shaojie Zhang, University of Central Florida, USA
Shiaofen Fang, Indiana University-Purdue University Indianapolis, USA
Shuaicheng Li, City University of Hong Kong, China
Shuang Luan, University of New Mexico, USA

Tao Huang, Chinese Academy of Sciences, China
Tianle Ma, Oakland University, USA
Ting Hu, Queen's University, Canada
Travis Johnson, The Ohio State University, USA
Wanding Zhou, Children's Hospital of Philadelphia, USA
Wei Zhang, University of Central Florida, USA
Weichun Huang, National Institutes of Health, USA
Jia Wen, University of North Carolina at Chapel Hill, USA
Xiaofeng Song, Nanjing University of Aeronautics & Astronautics, China
Xiaohui Yao, Harbin Institute of Technology, China
Xiaoming Liu, University of South Florida, USA
Xiaowen Liu, Tulane University, USA
Yang Shen, Texas A&M University, USA
Yaping Liu, Cincinnati Children's Hospital Medical Center, USA
Yonghui Wu, University of Florida, USA
Yongsheng Bai, University of Michigan, USA
Hui Yu, University of New Mexico, USA
Yu Xue, Huazhong University of Science and Technology, China
Yuan Luo, Northwestern University, USA
Yufei Huang, University of Texas at San Antonio, USA
Yufeng Shen, Columbia University, USA
Yufeng Wang, University of Texas at San Antonio, USA
Yulin Dai, University of Texas Health Science Center at Houston, USA
Yunlong Liu, Indiana University-Purdue University Indianapolis, USA
Yunyun Zhou, Children's Hospital of Philadelphia, USA
Zechen Chong The University of Alabama at Birmingham, USA
Zhandong Liu, Baylor College of Medicine, USA
Zhe He, Florida State University, USA
Zheng Wang, University of Miami, USA
Zhengqing Ouyang, University of Massachusetts at Amherst, USA
Zhifu Sun, Mayo Clinic, USA
Zuoyi Zhang, Indiana University, USA
Ruifeng Huang, Harvard University, USA
Yu-Ping Wang, Tulane University, USA
Aman Kaushik, Jiangnan University, China
Jie Hao, University of Pennsylvania, USA
Jingcheng Du, UT Health Science Center at Houston, USA
Cong Liu, Columbia University, USA
Hongchang Gao Temple University, USA

Publication Committee

Yan Guo, Co-Chair, University of New Mexico, USA
Wei Zhang, Co-Chair, University of Central Florida, USA

Workshop/Tutorial Committee

Kin Fai Au, Co-Chair, The Ohio State University, USA
Yulin Dai, Co-Chair, The University of Texas Health Science Center at Houston, USA

Publicity Committee

Yu Xue, Co-Chair, Huazhong University of Science and Technology, China
Kwangsik Nho, Co-Chair, Indiana University, USA

Award Committee

Jinchuan Xing, Co-Chair, Rutgers University, USA
Lanjing Zhang, Co-Chair, Princeton Medical Center, USA
Yong Cheng, Co-Chair, St. Jude Children’s Research Hospital, USA

Trainee Committee

James Havrilla, Chair, Children's Hospital of Philadelphia

Local Organization Committee

Dokyoon Kim, Co-Chair, University of Pennsylvania, USA
Wanding Zhou, Co-Chair, Children's Hospital of Philadelphia, USA

Website Chairs

Junjie Chen, Temple University, USA
Bin Li, Temple University, USA
Kai Wang, Children's Hospital of Philadelphia, USA

International Conference on Intelligent Biology and Medicine (ICIBM 2021) Program At-a-Glance (August 08-10, 2021, virtual via Zoom, US Eastern Time)

Sunday, August 8th, 2021

WORKSHOPS AND TUTORIALS			
9:00 - 11:30am Session Chair: Kin-Fai Au	Dr. Sudhir Kumar and team Temple University **A Live Tutorial on Molecular Evolutionary Genetics Analysis (MEGA)**	10:00 - 11:30am Session Chair: Yulin Dai	Drs. Jiaxin Fan and Rui Xiao The University of Pennsylvania Perelman School of Medicine **Statistical Methods for Allele-Specific Expression Analysis Using RNA Sequencing Data**
11:30am - 2:00pm	***Break***		
2:00 - 4:30pm Session Chair: Kin-Fai Au	Drs. Yulin Dai and Hyun-Hwan Jeong The University of Texas Health Science Center at Houston **Advanced Computational Analyses of Single-Cell RNA Sequencing Data**	3:00 - 4:30pm Session Chair: Kin-Fai Au	Drs. Scott Williams, Weihua Guan, Steven F. Jennings, Philip R.O. Payne, and Jonathan Stubblefield Case Western Reserve University; The University of Minnesota; University of Arkansas at Little Rock; Washington University in St. Louis; Arkansas State University **No Boundary Thinking in Bioinformatics**
4:30pm	***Adjourn***		

Monday, August 9th, 2021 (US Eastern Time)

8:15-8:30am	**Opening Remarks (Mindy Shi, Li Shen)**
8:30-9:10am Session Chair: Li Shen	**Keynote Lecture** **James S. Duncan, Ph.D.** **Fellow of IEEE and AIMBE, Fellow and former President of MICCAI society.** **Ebenezer K. Hunt Professor of Electrical Engineering & Radiology and Biomedical Imaging** **Yale University, USA** Title: *Neuroimage Analysis in Autism: from Model-Based Estimation to Data-driven Learning*
9:10 -9:30am Session Chair: Mindy Shi	**Eminent Scholar Talk** **Feng Yue, Ph.D.** **Director, Center for Cancer Genomics, Robert H. Lurie Comprehensive Cancer Center** **Director, Institute for Augmented Intelligence in Medicine - Center for Advanced Molecular Analysis** **Duane and Susan Burnham Professor of Molecular Medicine** **Associate Professor, Department of Biochemistry and Molecular Genetics** **Feinberg School of Medicine, Northwestern University, USA** Title: *Genome-Wide Detection and Functional Characterization of Enhancer Hijacking in Cancer Genomes*
9:30-9:45am	***Break***
PAPER PRESENTATION SESSIONS	
I. Genomics and Networks Session Chairs: Li Liu, Jim Havrilla	
9:45-10:00am	*pHisPred: A Tool for the Identification of Histidine Phosphorylation Sites by Integrating Amino Acid Patterns and Properties*

	Jian Zhao, Minhui Zhuang, Jingjing Liu, Meng Zhang, Cong Zeng, Bin Jiang, Jing Wu, Xiaofeng Song
10:00-10:15am	*Mining Functional Gene Modules by Multi-View NMF of Phenome-Genome Association* Xu Jin, WenQian He, MingMing Liu, Lin Wang, YaoGong Zhang, YingJie Xu, Ling Ma, YaLou Huang, MaoQiang Xie
10:15-10:30am	*Signaling Interaction Link Prediction Using Deep Graph Neural Networks Integrating Protein-Protein Interactions and Omics Data* Jiarui Feng, Amanda Zeng, Yixin Chen, Philip Payne, Fuhai Li
10:30-10:45am	*eSMC: A Statistical Model to Infer Admixture Events from Individual Genomics Data* Yonghui Wang, Zicheng Zhao, Xinyao Miao, Yinan Wang, Xiaobo Qian, Lingxi Chen, Changfa Wang, Shuaicheng Li
10:45-11:00am	*Identifying Genes Associated with Brain Bolumetric Differences Through Tissue Specific Transcriptomic Inference from GWAS Summary Data* Hung Mai, Jingxuan Bao, Paul M. Thompson, Dokyoon Kim, Li Shen
11:00-11:15am	***Break***
11:15-11:30am	*SCSilicon: a Tool for Synthetic Single-Cell DNA Sequencing Data Generation* Xikang Feng, Lingxi Chen
11:30-11:45am	*McSNAC: A Software to Approximate First-Order Signaling Networks from Mass Cytometry Data* Darren Wethington, Sayak Mukherjee, Jayajit Das
11:45am-12:00pm	*FSF-GA: A Feature Selection Framework for Phenotype Prediction Using Genetic Algorithms* Mohammad Erfan Mowlaei, Xinghua Shi
12:00-12:15pm	*Deciphering the Role of RNA Structure in Translation Efficiency* Jianan Lin, Yang Chen1, Yuping Zhang, Haifan Lin, Zhengqing Ouyang
12:15-12:30pm	*Integrative Analysis of Summary Data from GWAS and eQTL Studies Implicates Genes Differentially Expressed in Alzheimer's Disease* Brian Lee, Xiaohui Yao, Li Shen
12:30-2:30pm	**Break**

II. Deep Learning

Session Chairs: Cong Liu, Lanjing Zhang

2:30-2:45pm	*CAISC: A Software to Integrate Copy Number Variation and Single Nucleotide Mutations for Genetic Heterogeneity Profiling and Subclone Detection by Single-cell RNA Sequencing* Jeerthi Kannan, Liza Mathews, Zhjie Wu, Neal S Young, Shouguo Gao
2:45-3:00pm	*The Versatile Alignment Tool (VAT): A High-Performance Multi-Purpose Short Sequence Mapping Toolkit* Cuncong Zhong, Xiangtao Liu
3:00-3:15pm	*Neural Representations of Cryo-EM Maps and a Graph-Based Interpretation* Nathan Ranno, Dong Si
3:15-3:30pm	*LENRM: A New Noise Reduction Method Based on Local Expansion for Detecting Overlapping Protein Complexes* Lei Xue, Xu Qing Tang
3:30-3:45pm	*CrisprVi: A Software for Visualizing and Analyzing CRISPR Sequences of Prokaryotes* Lei Sun, Fu Yan, Gongming Wang, Jinbiao Wang, Yun Li, Jinlin Huang
3:45-4:00pm	**Break**
4:00-4:15pm	*Deep Multiview Learning to Identify Imaging-driven Subtypes in Mild Cognitive Impairment* Yixue Feng, Mansu Kim, Xiaohui Yao, Kefei Liu, Qi Long, Li Shen for the Alzheimer's Disease Neuroimaging Initiative
4:15-4:30pm	*B-assembler: A Circular Bacterial Genome Assembler* Fengyuan Huang, Li Xiao, Min Gao, Ethan J. Vallely, Kevin Dybvig, Prescott Atkinson, Ken B. Waites, Zechen Chong
4:30-4:45pm	*DISTEMA: Distance Map-Based Estimation of Single Protein Model Accuracy with Attentive 2D Convolutional Neural Network* Xiao Chen, Jianling Cheng
4:45-5:00pm	*BVMHC: Bilateral and Variable Long Short Term Memory Networks Based Major Histocompatibility Complex Binding Prediction*

	Limin Jiang, Hui Yu, Jijun Tang, Fei Guo, Yan Guo
5:00-5:15pm	*Integrative Analysis of eQTL and GWAS Summary Statistics Reveals Transcriptomic Alteration in Alzheimer Brains* Pradeep Varathan, Priyanka Gorijala, Tanner Jacobson, Danai Chasioti, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Jingwen Yan
5:15-5:45pm Session Chair: Jim Havrilla	***Flash Talk – Machine Learning and Bioinformatics*** 1. *Contribution of Transposable Elements to Tissue-Specific Gene Regulation in Human* Arsala Ali, Ping Liang 2. *Analysis of Factors Impacting the Quality of de novo Genome Assemblies.* Haimeng (Jerry) Tang, Ping Liang, Adonis Skandalis, Miriam Richards 3. *DRAGOM: Classification and Quantification of Noncoding RNA in Metagenomic Data* Ben Liu, Sirisha Thippabhotla, Jun Zhang, Cuncong Zhong 4. *iMPP: Integrated De Novo Gene Prediction for Metagenomics Data* Sirisha Thippabhotla, Ben Liu, Cuncong Zhong 5. *Bayesian Mixed-Effect Higher-Order Hidden Markov Models with Applications to Predictive Healthcare Using Electronic Health Record* Ying Liao, Yisha Xiang, Zhigen Zhao, Di Ai 6. *Gene Regulatory Networks and Biomarkers Jointly Inferred by Genome-Wide Genetic, Epigenetic, and Regulatory Factors in Multiple Sclerosis* Astrid M Manuel, Yulin Dai, Hyun-Hwan Jeong, Zhongming Zhao 7. *Charting the Proteome Landscape in Major Psychiatric Disorders: From Biomarkers to Biological Pathways* Brisa S. Fernandes, Yulin Dai, Peilin Jia, Zhongming Zhao 8. *A Graph Neural Network Model to Estimate Cell-Wise Metabolic Flux Using Single Cell RNA-seq Data* Wennan Chang,Norah Alghamdi, Pengtao Dang, Xiaoyu Lu, Changlin Wan, Silpa Gampala, Yong Zang, Melissa Fishel, Sha Cao, Chi Zhang
5:45-8:00pm	**Break**

8:00 -8:40pm Session Chair: Kai Wang	**Keynote Lecture** **Chunhua Weng, Ph.D., FACMI, FIAHSI** **Professor of Biomedical Informatics, Member of Data Science Institute** **Co-Director for Biomedical Informatics, Irving Institute for Clinical and Translational Research** **Columbia University Irving Medical Center, New York, NY, USA** Title: *Augmented Intelligence for Clinical Trials: from Participant Selection to Evidence Appraisal*
8:40 -9:00pm Session Chair: Li Shen	**Eminent Scholar Talk** **Graciela Gonzalez-Hernandez, Ph.D.** **Associate Professor of Informatics** **The Perelman School of Medicine, University of Pennsylvania, USA** Title: *Challenges in Digital Epidemiology: Using Social Media Mining for Health Research*
9:00-9:15pm	***Break***
PAPER PRESENTATION SESSIONS	
III. Medical Informatics and Decision Making Session Chairs: Yunyun Zhou, Tianle Ma	
9:15-9:30pm	*AEDNav: Indoor Navigation for Locating Automated External Defibrillator* Gaurav Rao, Vijay Mago, Pawan Lingras, David W. Savage
9:30-9:45pm	*Estimating the Optimal Linear Combination of Predictors Using Spherically Constrained Optimization* Priyam Das, Debsurya De, Raju Maiti, Mona Kamal, Katherine A. Hutcheson, Clifton D. Fuller, Bibhas Chakraborty, Christine B. Peterson
9:45-10:00pm	*DENSEN: a Convolutional Neural Network for Estimating Chronological Ages from Panoramic Radiographs* Xuedong Wang, Xinyao Miao, Yanle Liu, Yin Chen, Xiao Cao, Yuchen Zhang, ShuaichengLi and

	Qin Zhou
10:00-10:15pm	*Identification of Multimodal Brain Imaging Association via A Parameter Decomposition based Sparse Multi-view Canonical Correlation Analysis Method* Jin Zhang, Huiai Wang, Ying Zhao, Lei Guo, Lei Du, the Alzheimer's Disease Neuroimaging Initiative
10:15-10:30pm	*An Integrated Interactive COVID-19 Dashboard for Individual Risk Analysis and Real-time Trend Analysis* Josh Voytek, Maria Maltepes, Anna Lengner, Jay S. Patel, Huanmei Wu
10:30-10:45pm	*Mining Comorbidities of Opioid Use Disorder from FDA Adverse Event Reporting System and Patient Electronic Health Records* Yiheng Pan, Rong Xu
10:45-11:00pm	**Break**
11:00-11:15pm	*PheNominal: An EHR-Integrated Web Application for Structured Deep Phenotyping at the Point of Care* James M. Havrilla, Anbumalar Singaravelu, Dennis M. Driscoll, Leonard Minkovsky, Ingo Helbig, Livija Medne, Kai Wang, Ian Krantz, Bimal R. Desai
11:15-11:30pm	*Natural Language Processing to Identify Lupus Nephritis Phenotype in Electronic Health Records* Yu Deng, Jennifer A. Pacheco, Anh Chung, Chengsheng Mao, Joshua C. Smith, Juan Zhao, Wei-Qi Wei, April Barnado, Chunhua Weng, Cong Liu, Adam Cordon, Jingzhi Yu, Yacob Tedla, Abel Kho, Rosalind Ramsey- Goldman, Theresa Walunas, Yuan Luo
11:30-11:45pm	*Expediting Knowledge Acquisition by a Web Framework for Knowledge Graph Exploration and Visualization (KGEV): a Case Study on COVID-19* Jacqueline Peng, David Xu, Ryan Lee, Siwei Xu, Yunyun Zhou, Kai Wang
11:45pm-12:00am	*Disparities in Social Determinants among Performances of Mortality Prediction with Machine Learning for Sepsis Patients* Hanyin Wang, Yikuan Li, Andrew Naidech, and Yuan Luo
12:00-12:15am	*On the Role of Deep Learning Model Complexity in Adversarial Robustness for Medical Images* David Rodriguez, Tapsya Nayak, Yidong Chen, Ram Krishnan, Yufei Huang

12:15-12:30am	*Application of Unsupervised Deep Learning Algorithms for Identification of Specific Clusters of Chronic Cough Patients from EMR Data* Wei Shao, Xiao Luo, Zuoyi Zhang, Zhi Han, Vasu Chandrasekaran, Vladimir Turzhitsky, Vishal Bali, Anna R. Roberts, Megan Metzger, Jarod Baker, Carmen La Rosa, Jessica Weaver, Paul Dexter, Kun Huang
12:30am	**Adjourn**

Tuesday, August 10th, 2021 (US Eastern Time)

8:25-8:30am	**Introduction (Mindy Shi, Li Shen)**
8:30-9:10am Session Chair: Mindy Shi	**Keynote Lecture** **Ben Raphael, Ph.D.** **Professor of Computer Science** **Princeton University, USA** Title: *Quantifying Tumor Heterogeneity across Time and Space*
9:10-9:30am Session Chair: Kai Wang	**Eminent Scholar Talk** **Kai Tan, Ph.D.** **Professor of Pediatrics** **Children's Hospital of Philadelphia** **Perelman School of Medicine** **The University of Pennsylvania, USA** Title: *Leverage Systems Biology and Single-Cell Analysis to Discover Novel Therapeutic Targets*
9:30-9:45am	***Break***
PAPER PRESENTATION SESSIONS	
IV. Genomics, RNA Biology and Diseases Session Chairs: Jinchuan Xing, Yong Cheng	
9:45-10:00am	*Identifying Genetic Markers Enriched by Brain Imaging Endophenotypes in Alzheimer's Disease* Mansu Kim, Ruiming Wu, Xiaohui Yao, Andrew J. Saykin, Jason H. Moore, Li Shen for the Alzheimer's Disease Neuroimaging Initiative
10:00-10:15am	*AutoCoV: Tracking the Early Spread of COVID-19 in Terms of the Spatial and Temporal Dynamics from Embedding Space by K-mer Based Deep Learning* Inyoung Sung, Sangseon Lee, Minwoo Pak, Yunyol Shin, Sun Kim

10:15- 10:30am	*A Framework to Trace Microbial Engraftment at the Strain Level During FMT* Yiqi Jiang, Shuai Wang, Xianglilan Zhang, Shuaicheng Li
10:30- 10:45am	*Exploration of Chemical Space with Partial Labeled Noisy Student Self-Training and Self-Supervised Graph Embedding: Application to Drug Metabolism* Yang Liu, Hansaim Lim, Lei Xie
10:45- 11:00am	*Novel lincRNA Discovery and Tissue-Specific Gene Expression across 30 Normal Human Tissues* Xianfeng Chen, Zhifu Sun
11:00- 11:15am	***Break***
11:15- 11:30am	*Bioinformatics Analysis Revealed that Functional Important miRNA Targeted Genes are Associated with Child Obesity Trait in Genome-wide Association Studies* Melinda Song, Jiaqi Yu, Binze Li, Julian Dong, Jeslyn Gao, Lulu Shang, Xiang Zhou, Yongsheng Bai
11:30- 11:45am	*MSPCD: Predicting circRNA-disease associations via integrating multi-source data and hierarchical neural network* Lei Deng, Dayun Liu, Yizhan Li, Runqi Wang, Junyi Liu, Jiaxuan Zhang, Hui Liu
11:45am - 12:00pm	*Gene Co-Expression Changes Underlying the Functional Connectomic Alterations in Alzheimer's Disease* Bing He, Priyanka Gorijala, Linhui Xie, Sha Cao, Jingwen Yan
12:00- 12:15pm	*Prioritization of Risk Genes in Multiple Sclerosis by a Refined Bayesian Framework Followed by Tissue-Specificity and Cell Type Feature Assessment* Andi Liu, Astrid M Manual, Yulin Dai, Zhongming Zhao
12:15- 12:30pm	*SARS-COV-2 as Potential microRNA Sponge in COVID-19 Patients* Chang Li, Rebecca Wang, Aurora Wu, Tina Yuan, Kevin Song, Yongsheng Bai, Xiaoming Liu
12:30- 2:30pm	***Break***
V. Cancer Informatics Session Chairs: Wei Zhang, Zhifu Sun	

2:30-2:45pm	*Bi-EB: An Empirical Bayesian Biclustering Algorithm for shared omics patterns between breast cancer tumor samples and breast cancer cell lines* Aida Yazdanparast, Lang Li, Chi Zhang, Lijun Cheng
2:45-3:00pm	*Constrained Tensor Factorization for Computational Phenotyping and Mortality Prediction in Patients with Cancer* Francisco Cai, Chengsheng Mao, Yuan Luo
3:00-3:15pm	*Small Molecule Modulation of Microbiota: A Systems Pharmacology Perspective* Qiao Liu, Bohyun Lee, Lei Xie
3:15-3:30pm	*Cost-effective Low-coverage Whole Genome Sequencing Assay for Glioma Risk Stratification* Jin Fu, Yingfeng Zhu, Xiaofeng Li, Jiajun Qin, Zhongrong Chen, Ziliang Qian, Jiping Sun, Xianzhen Chen
3:30-3:45pm	*The Profiling of Gut Microbiota During Colorectal Cancer Development* Jingjing Liu, Wei Dong, Jian Zhao, Jing Wu, Jinqiang Xia, Shaofei Xie, Xiaofeng Song
3:45-4:00pm	***Break***
4:00-4:15pm	*Model Performance and Interpretability of Semi-Supervised Generative Adversarial Networks to Predict Oncogenic Variants with Unlabeled Data* Zilin Ren, Quan Li , Kajia Cao , Marilyn M Li , Yunyun Zhou, Kai Wang
4:15-4:30pm	*Copy Number Variation of Urine Exfoliated Cells s by Low-Coverage Whole Genome Sequencing for Diagnosis of Prostate Adenocarcinoma: A Prospective Cohort Study* Youyan Guan, Xiaobing Wang, Kaopeng Guan, Dong Wang, Xingang Bi, Zhendong Xiao, Zejun Xiao, Xingli Shan, Linjun Hu, Jianhui Ma, Changling Li, Yong Zhang, Jianzhong Shou, Baiyun Wang, Ziliang Qian, Nianzeng Xing
4:30-4:45pm	*Association of the Tissue Infiltrated and Peripheral Blood Immune Cell Subsets with Response to Radiotherapy for Rectal Cancer* Xueling Li, Min Zhu, Xingjie Li, Xu Cheng, Xingxu Yi, Fang Ye, Xiaolai Li, Zongtao Hu, Liwei Zhang, Jinfu Nie
4:45-5:00pm	*Comparison of Four Supervised Feature Selection Algorithms Leading to Top features and Gene*

	Signatures from Multi-Omics Data in Cancer Tapas Bhadra, Saurav Mallik, Neaj Hasan, Zhongming Zhao
5:00-5:15pm	*A Landscape of Immune Cell Types in Tumor Microenvironment Associated with Prognosis and Sensitivity of Radiotherapy* Xingjie Li, Min Zhu, Xueling Li
5:15-5:45pm Session Chair: Jim Havrilla	***Flash Talk – Single Cell Omics and Disease Informatics*** *9. Converting Tabular Data into Images for Anti-Cancer Drug Response Prediction Using Convolutional Neural Networks* Yitan Zhu, Thomas Brettin, Fangfang Xia, Alexander Partin, Maulik Shukla, Hyunseung Yoo, Yvonne A. Evrard, James H. Doroshow, Rick L. Stevens *10. Functional Screening of 3'-UTR Variants Combined with Genome-wide Association Identifies Causal Regulatory Mechanisms Impacting Alcohol Consumption* Andy B Chen, Kriti S. Thapa, Hongyu Gao, Jill L Reiter, Hongmei Gu, Junjie Zhang, Xiaoling Xuei, Dongbing Lai, Yue Wang, Howard J. Edenberg, Yunlong Liu *11. A Machine-Learning Classifier for Predicting Aneuploidy Risk in Female IVF Patients* Siqi Sun, Maximilian Miller, Yanran Wang, Katarzyna M. Tyc, Richard T. Scott, Jr., Xin Tao, Yana Bromberg, Karen Schindler, Jinchuan Xing *12. Single Cell-Based Deconvolution of Liver Diseases Reveals γδ2 T Cells as a Marker in Hepatocellular Carcinoma Development* Rama Shankar, Mingdian Tan, Joseph W. Zagorski, Austin J. Goodyke, Jeremy Haskins, Shreya Paithankar, Dave Chesla, Samuel So, Mei-Sze Chua, Bin Chen *13. Brain Dynamic Functional Connectivity Predicts Treatment Response to Electroconvulsive Therapy in Major Depressive Disorder* Mohammad. S. E. Sendi,, Hossein Dini, Christopher C. Abbott, Vince D Calhoun *14. IPDB: Integrated Pregnancy Database with clinical and omics data* Parth G Kothiya, Huanmei Wu, David M. Haas, Shelley D. Dowden, Bobbie N Ray, Sara K Quinney *15. Therapeutic Re-Positioning of Amiloride: From Anti- hypertension to Anti-Cancer* Aleshia Seaton-Terry, Venkataswarup Tiriveedhi *16. Delineating Cell State Heterogeneity in Bladder Cancer* Antara Biswas, Sivasomasundari Arunarasu, Subhajyoti De

5:45-8:00pm	**Break**

8:00-8:40pm Session Chair: Zhongming Zhao	**Keynote Lecture** **Ying Xu, Ph.D.** **Fellow of AAAS** **Regents-GRA Eminent Scholar Chair** **Professor of Bioinformatics and Computational Biology** **University of Georgia, USA** Title: *Towards Developing a New Theory of Cancer Evolution*
8:40-9:00pm Session Chair: Kong Chen	**Eminent Scholar Talk** **Wei Chen, Ph.D.** **Professor of Pediatrics, Biostatistics, Human Genetics** **Director of Statistical Genetics Core** **University of Pittsburgh** **UPMC Children's Hospital of Pittsburgh** Title: *Understanding Age-related Macular Degeneration in the Era of Big Data*
9:00-9:15pm	***Break***
PAPER PRESENTATION SESSIONS	
VI. Transcriptomics Session Chairs: Rui Xiao, Xiaoming Liu	
9:15-9:30pm	*Revealing the Novel Complexity of Plant Long Non-Coding RNA by Strand-Specific and Whole*

	Transcriptome Sequencing for Evolutionarily Representative Plant Species Yan Zhu, Longxian Chen, Xiangna Hong, Han Shi, Xuan Li
9:30-9:45pm	*Identifying Alzheimer’s Genes via Brain Transcriptome Mapping* Jae Young Baik, Mansu Kim, Jingxuan Bao, Qi Long, Li Shen
9:45-10:00pm	*On Triangular Inequalities of Correlation-Based Distances for Gene Expression Profiles* Jiaxing Chen, Yen Kaow Ng, Lu Lin, Xianglilan Zhang, Shuaicheng Li
10:00-10:15pm	*Mouse Blood Cells Types and Aging Prediction using Penalized Latent Dirichlet Allocation* Xiaotian Wu, Yee Voan Teo, Nicola Neretti, Zhijin Wu
10:15-10:30pm	*Rewired Pathways and Disrupted Pathway Crosstalk in Schizophrenia Transcriptomes by Multiple Differential Coexpression Methods* Hui Yu, Yan Guo, Jingchun Chen, Xiangning Chen, Peilin Jia, Zhongming Zhao
10:30-10:45pm	***Break***
10:45-11:00pm	*kESVR: An Ensemble Model for Drug Response Prediction in Precision Medicine Using Cancer Cell Lines Gene Expression* Abhishek Majumdar,Yueze Liu,Yaoqin Lu, Shaofeng Wu, Lijun Cheng
11:00-11:15pm	*Revealing the Hadal Viral Community in the Sediment of New Britain Trench* Hui Zhou, Ping Chen, Mengjie Zhang, Jiawang Chen, Jiasong Fang, Xuan Li
11:15-11:30pm	*Alternative Splicing Induces Sample-Level Variation in Gene-Gene Correlations* Yihao Lu, Brandon L. Pierce, Pei Wang, Fan Yang, Lin S. Chen
11:30-11:45pm	*APA-Scan: Detection and Visualization of 3'-UTR APA with RNA-seq and 3'-end-seq Data* Naima Ahmed Fahmi, Khandakar Tanvir Ahmed, Jae-Woong Chang, Heba Nassereddeen, DeliangFan, Jeongsik Yong, Wei Zhang
11:45pm- 11:55pm Session Chair: Jinchuan Xing	***Award Ceremony***

11:55pm-12:05am Session Chairs: Mindy Shi, Li Shen	***Wrap-up and Closing Remarks***

Keynote Speaker
James S. Duncan, PhD
Monday, August 9, 2021
8:30 – 9:10 AM (US EDT)

BIO

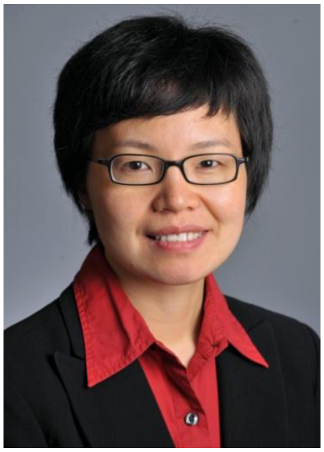
James S. Duncan is the Ebenezer K. Hunt Professor of Biomedical Engineering and a Professor of Radiology & Biomedical Engineering, Electrical Engineering and Statistics & Data Science at Yale University. Dr. Duncan received his B.S.E.E. with honors from Lafayette College (1973), and his M.S. (1975) and Ph.D. (1982) both in Electrical Engineering from the University of California, Los Angeles. Dr. Duncan has been at Yale since 1983. He has served as the Acting Chair and is currently Director of Undergraduate Studies for Biomedical Engineering. Dr. Duncan's research efforts have been in the areas of computer vision, image processing, and medical imaging, with an emphasis on biomedical image analysis and image-based machine learning. He has published over 300 peer-reviewed articles in these areas and has been the principal investigator on a significant number of peer-reviewed grants from both the National Institutes of Health and the National Science Foundation over the past 35 years. He is a Life Fellow of the Institute of Electrical and Electronic Engineers (IEEE), and a Fellow of the American Institute for Medical and Biological Engineering (AIMBE) and of the Medical Image Computing and Computer Assisted Intervention (MICCAI) Society. In 2014 he was elected to the Connecticut Academy of Science & Engineering. He has served as co-Editor-in-Chief of Medical Image Analysis, as an Associate Editor of IEEE Transactions on Medical Imaging, and on the editorial boards of Pattern Analysis and Applications, the Journal of Mathematical Imaging and Vision, "Modeling in Physiology" of The American Physiological Society and the Proceedings of the IEEE. He is a past President of the MICCAI Society. In 2012, he was elected to the Council of Distinguished Investigators, Academy of Radiology Research and in 2017 received the "Enduring Impact Award" from the MICCAI Society.

Keynote Speaker
James S. Duncan, PhD
Monday, August 9, 2021
8:30 – 9:10 AM (US EDT)

Title: Neuroimage Analysis in Autism: from Model-Based Estimation to Data-driven Learning

Abstract: Functional magnetic resonance imaging (fMRI) has been shown to be helpful for the study of autism spectrum disorders (ASD). This talk will describe the evolution of efforts in this area within our group that carry promise for producing objective biomarkers for ASD, as well as predicting patient response to a behavioral therapy known as Pivotal Response Treatment (PRT), using task-based fMRI. Such biomarkers would provide an important step in better understanding the underlying pathophysiology of ASD that could help with objective and personalized diagnosis, provide new targets for development of new treatments, and provide a way to monitor patient progress. Initially a robust, group-wise unified Bayesian framework to detect both hyper and hypo-active communities from connectivity maps will be described. Next, more recent work will be presented that has focused on deriving ASD biomarkers from individual subject's time-series data, based on the classification of individual subjects (into ASD or typical control) and identifying spatially-specific key regions using convolutional neural networks and ablation analysis of regions. Finally, a strategy based on recurrent neural networks (using long-short-term memories or LSTMs) will be presented that predicts patient response to PRT behavioral therapy from baseline imaging while incorporating subject-specific phenotypic information for network initialization.

Keynote Speaker
Chunhua Weng, PhD
Monday, August 9, 2021
8:00 – 8:40 PM (US EDT)



BIO

Dr. Chunhua Weng is a tenured Full Professor of Biomedical Informatics at Columbia University and an elected fellow of both American College of Medical Informatics (ACMI) and International Academy of Health Sciences Informatics (IAHSI). She has been co-leading the Biomedical Informatics Resource for the Columbia CTSA (The Irving Institute for Clinical and Translational Science) since 2011. She is also an Associate Editor for Journal of Biomedical Informatics. Dr. Weng holds a Ph.D. in Biomedical and Health Informatics from University of Washington at Seattle. As an active researcher in the field of Clinical Research Informatics since 2000, Dr. Weng has published extensively on data-driven optimization of clinical trial eligibility criteria, scalable and portable electronic phenotyping, electronic health records (EHR) data quality assessment and data analytics, and text knowledge engineering using a variety of text (e.g., EHR narratives, PubMed abstracts and clinical trial summaries).

Keynote Speaker
Chunhua Weng, PhD
Monday, August 9, 2021
8:00 – 8:40 PM (US EDT)

Title: Augmented Intelligence for Clinical Trials: from Participant Selection to Evidence Appraisal

Abstract: Clinical trials are the foundation for the advances in medicine. However, they are often criticized for their high expenses, lengthy study time, and poor results generalizability. In the life cycle of clinical trials, from study design to conduct, and to evidence comprehension and synthesis, there are many unmet user needs among clinical researchers, patients, and clinicians, causing suboptimal decisions and potentially compromised clinical trials. In this talk, I will present recent research in my lab that aims to provide Augmented Intelligence to these different stakeholders of clinical trials in order to augment their decision making. I will particularly focus on natural language processing and data-driven methods for optimizing clinical trials participant selection and for automating evidence extraction and appraisal.

Keynote Speaker
Ben Raphael, PhD
Tuesday, August 10, 2021
8:30 – 9:10 AM (US EDT)

BIO

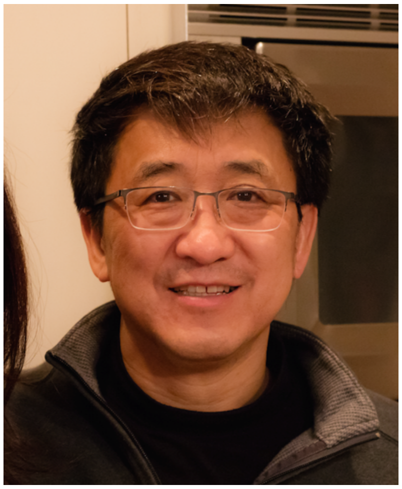
Ben Raphael is a Professor of Computer Science at Princeton University. His research focuses on the design of combinatorial and statistical algorithms for the interpretation of biological data. Recent areas of emphasis include cancer evolution, network/pathway analysis of germline and somatic mutations, single-cell and spatial DNA/RNA sequencing, and structural variation in human and cancer genomes. His group's algorithms have been used in multiple projects from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). He received an S.B. in Mathematics from MIT, a Ph.D. in Mathematics from the University of California, San Diego (UCSD), completed postdoctoral training in Bioinformatics and Computer Science at UCSD, and was on the faculty of Brown University (2006-2016). He is a recipient of the 2021 Innovator Award from the International Society for Computational Biology, the Alfred P. Sloan Research Fellowship, the NSF CAREER award, and a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. He is elected Fellow of the International Society for Computational Biology (2020).

Keynote Speaker
Ben Raphael, PhD
Tuesday, August 10, 2021
8:30 – 9:10 AM (US EDT)

Title: Quantifying Tumor Heterogeneity across Time and Space

Abstract: Tumors are heterogeneous mixtures of normal and cancerous cells with distinct genetic and transcriptional profiles. In this talk, I will present several computational approaches to quantify tumor heterogeneity and to study tumor evolution using single-cell and spatial sequencing technologies. For single-cell DNA sequencing data, I will describe algorithms to reconstruct tumor evolution from multiple types of somatic mutations and will use these approaches to analyze changes in tumor genomes over time. For spatial transcriptomics data, I will introduce algorithms to detect genomic aberrations and to align and integrate data from multiple adjacent tissue sections leveraging both spatial and transcriptional similarity. I will illustrate applications of these methods to quantify spatial heterogeneity in several cancer types.

Keynote Speaker
Ying Xu, PhD
Tuesday, August 10, 2021
8:00 – 8:40 PM (US EDT)



BIO

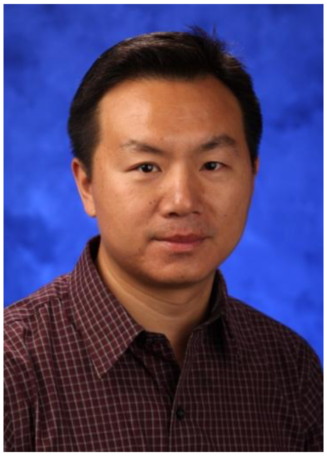
Ying Xu has been the "Regents and Georgia Research Alliance Eminent Scholar" Chair of bioinformatics and computational biology and Professor in Biochemistry and Molecular Biology Department since 2003 and was the Founding Director of the Institute of Bioinformatics, (2003 – 2011) the University of Georgia (UGA). He received his Ph.D. degree in theoretical computer science from the University of Colorado at Boulder in 1991. He started his bioinformatics career in 1993 to take part in the Human Genome Project when he joined Oak Ridge National Lab, where he worked for ten and half years. In 2003, he was recruited to the UGA to build the Institute of Bioinformatics. His current research interests are in cancer bioinformatics and systems biology, particularly in cancer metabolism. He has over 300 publications, including five books, with total citations more than 16,000 and H-Index = 66; and has given over 250 invited/contributed talks at conferences, research organizations and universities. He is a fellow of the AAAS (2007) and served as the Editor-in-Chief of IEEE/ACM Transaction in Computational Biology and Bioinformatics between Jan 2013 and Dec. 2016.

Keynote Speaker
Ying Xu, PhD
Tuesday, August 10, 2021
8:00 – 8:40 PM (US EDT)

Title: Towards Developing a New Theory of Cancer Evolution

Abstract: *Is it possible to derive the driving forces and associated mechanisms of cancer formation and evolution from the cancer omic data?* An evolutionary framework is needed to accomplish such ambitious goals. However, little can be borrowed from the existing cancer theories to guide such efforts since too many cancer-related questions cannot be addressed by them. In this presentation, I will outline a "stress-adaptation" framework, through which the many seemingly very unusual and counter-intuitive behaviors of cancer can be interpreted as the adaptative steps to the specific stressors that such cells uniquely encounter. Knowing that the cost to become cancerous is very high, say, consuming 20 to 30-fold more glucose than their matching normal cells and over 95% of them die soon after they are created, we expect that the cost for staying unchanged would be even higher. Our basic hypothesis is: **cancer is a survival process under such (to-be-identified) stress, in which cells MUST divide as otherwise they will die, 100%**. With this guiding hypothesis, we have discovered that many of the cancer biology questions have to be addressed at the basic chemistry (or physical) homeostasis level. Our data analyses and modeling have revealed that the affected cells must make fundamental changes in their cells' definition, for them to overcome the persistently disrupted homeostases, potentially the **cancer-defining stressors**, which are ultimately due to chronic inflammation and local iron accumulation. Specifically, they need to transform themselves from cells of a multi-cellular organism (i.e., a human being) to a unicellular "organism" to enable the key survival pathways, which is accomplished through selection of mutations in large numbers of genes of specific functions. A variety of cancer behaviors can be explained in terms of this transformation and associated reprogrammed metabolisms. This is an extremely exciting and challenging puzzle-solving problem; and we welcome interested people to join our effort to develop, together, a fundamentally novel theory of cancer evolution.

Eminent Scholar Talk
Feng Yue, PhD
Monday, August 9, 2021
9:10 – 9:30 AM (US EDT)



BIO

Dr. Feng Yue is the founding director of the Center for Cancer Genomics at the Robert H. Lurie Comprehensive Cancer Center of Northwestern University, director of the Center for Advanced Molecular Analysis at Northwestern Institute for Augmented Intelligence in Medicine, and the Duane and Susan Burnham Professor of Molecular Medicine. He is a tenured Associate Professor in the Department of Biochemistry and Molecular Genetics and the Department of Pathology.

The main research area for Dr. Yue's group is epigenomics and 3D genome organization in the context of human diseases. He has been a long-time member of several large NIH-funded consortia, such as the ENCODE Consortium, the Roadmap Epigenomics Project, and the 4D Nucleome Project. He led the overall organizing and analysis effort for the mouse ENCODE consortium. Currently, he serves as the co-chair for the Joint Analysis Working group in the 4D Nucleome Project, leading the consortium effort to integrate multiple data types to profile the 3D genome organization and its relationship with gene regulation. His group has a strong interested in cancer genomics and has demonstrated how epigenome and 3D genome structure are altered and led to gene dysregulation in different types of tumors, such as leukemia, bladder cancer, and brain tumor. More recently, his group and their collaborators show that Hi-C can be used as tool for systematic discovery of SVs in the genome and also reported widespread neo-TADs and enhancer hijacking events, which potentially contribute to gene misregulation in cancer cells. The long-term goal for Dr. Yue's research is to reveal the key cancer-specific or subtype-specific regulators and pathways, which can be potentially used as cancer biomarkers and therapeutic targets.

Eminent Scholar Talk
Feng Yue, PhD
Monday, August 9, 2021
9:10 – 9:30 AM (US EDT)

Title: Genome-Wide Detection and Functional Characterization of Enhancer Hijacking in Cancer Genomes

Abstract: Recent efforts have shown that structural variations (SVs) can disrupt three-dimensional genome organization and induce enhancer hijacking, yet no computational tools exist to identify such events from chromatin interaction data. Here, we develop NeoLoopFinder, a computational framework to identify the chromatin interactions induced by SVs, including interchromosomal translocations, large deletions and inversions. Our framework can automatically resolve complex SVs, reconstruct local Hi-C maps surrounding the breakpoints, normalize copy number variation and allele effects and predict chromatin loops induced by SVs. We applied NeoLoopFinder in Hi-C data from 50 cancer cell lines and primary tumors and identified tens of recurrent genes associated with enhancer hijacking. To experimentally validate NeoLoopFinder, we deleted the hijacked enhancers in prostate adenocarcinoma cells using CRISPR–Cas9, which significantly reduced expression of the target oncogene. In summary, NeoLoopFinder enables identification of critical oncogenic regulatory elements that can potentially reveal therapeutic targets.

Eminent Scholar Talk
Graciela Gonzalez-Hernandez, PhD
Monday, August 9, 2021
8:40 – 9:00 PM (US EDT)

BIO

Dr. Gonzalez Hernandez is a recognized expert and leader in natural language processing (NLP) applied to bioinformatics, medical/clinical informatics, and public-health informatics. After 11 years at the Department of Biomedical Informatics at Arizona State University, she joined the University of Pennsylvania and established the Health Language Processing Lab(link is external) within the Institute of Biomedical Informatics(link is external). Her recent work focuses on NLP applications for public-health monitoring and surveillance and is funded by R01 grants from the National Library of Medicine and the National Institute of Allergy and Infectious Diseases.

Her work on social media mining for pharmacovigilance has resulted in 10 publications in prestigious conferences and journals. Examples include work on ADR extraction in *the Journal of the American Medical Informatics Association (JAMIA)* and on prescription-drug abuse in *Drug Safety*. A *Journal of Biomedical Informatics* publication was selected as one of Elsevier/Atlas's 10 articles with greatest potential social impact (link is external), an honor among papers in more than 2500 journals. Her work in this area also caught the attention of the FDA, which awarded her a grant to develop these methods for monitoring nutritional supplements.

Her work on enriching geospatial information for phylogeography, in collaboration with Dr. Matthew Scotch, uses NLP for the automatic extraction of relevant geospatial data from the literature and for linkage to GenBank records. Preliminary work in this area resulted in publications in *JAMIA* and in *Oxford Bioinformatics*, and in and a presentation at ISMB in Dublin in 2015.

Dr Gonzalez served as a member of the NIH BLIRC panel from 2008 to 2013. She is a regular reviewer for a number of prestigious journals and conferences, including PLoS One, PLoS Computational Biology, *JAMIA* and *BMC Bioinformatics*. Her prior funding also included funding under the Arizona Alzheimer's Disease Center, a P30 NIA Center, as director of the Data Core from 2008 to 2016.

Eminent Scholar Talk
Graciela Gonzalez-Hernandez, PhD
Monday, August 9, 2021
8:40 – 9:00 PM (US EDT)

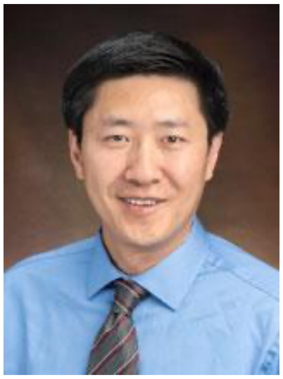
Title: Challenges in Digital Epidemiology: Using Social Media Mining for Health Research

Abstract: Social media has grown in popularity for health-related research as it has become evident that it can be a good source of patient insights. Be it Twitter, Reddit, Instagram, Facebook, Amazon reviews or health forums, researchers have collected and processed user comments and published countless papers on different uses of social media data. Using these data presents many challenges when it needs to be used in epidemiology. From identifying the right cohort and reducing bias to finding the 'needle in the haystack', social media data is sometimes misused and frowned upon when it is not properly handle. I will discuss some aspects of how solid scientific principles and careful design of natural language processing methods can help 'tame' the noise in social media data and enable digital epidemiology.

Some relevant publications:

- https://www.nature.com/articles/s41746-019-0170-5
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233076/
- https://doi.org/10.1093/jamia/ocz156
- https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2767638
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230947

Eminent Scholar Talk
Kai Tan, PhD
Tuesday, August 10, 2021
9:10 – 9:30 AM (US EDT)



BIO

Dr. Kai Tan is Professor of Pediatrics at the Children's Hospital of Philadelphia and the University of Pennsylvania Perelman School of Medicine. He is the co-leader of pediatric oncology program at the Abramson Cancer Center and director of the Center for Single Cell Biology at CHOP. His research uses systems biological approaches to studying transcriptional and epigenetic regulation of hematopoiesis and pediatric cancers.

Eminent Scholar Talk
Kai Tan, PhD
Tuesday, August 10, 2021
9:10 – 9:30 AM (US EDT)

Title: Leverage systems biology and single-cell analysis to discover novel therapeutic targets

Abstract: Therapy resistance is a major cause of death in cancer. This is due to the intrinsic genetic redundancy and heterogeneity of cancer cells. I will present computational methods that use multi-omics bulk and single-cell data to identify novel combinatorial therapeutic targets and rare cancer clones that drive oncogenesis and treatment response.

Eminent Scholar Talk
Wei Chen, PhD
Tuesday, August 10, 2021
8:40 – 9:00 PM (US EDT)

BIO

Dr. Wei Chen, PhD, is Professor of Pediatrics, Biostatistics, and Human Genetic in the School of Medicine, University of Pittsburgh, Director of Statistical Genetics Core at UPMC Children's Hospital of Pittsburgh. Dr. Chen's research interests include developing statistical and computational methods for high-throughput data, such as DNA sequencing, methylation, bulk and single cell RNA sequencing data, and retinal image, with applications to complex diseases, mainly on childhood asthma and age-related macular degeneration (AMD). He has served as Principal Investigator (PI) or co-PI on 11 grants from National Institute of Health (NIH), National Science Foundation (NSF), and foundations. His research activities have resulted in the publication of over 130 peer-reviewed papers including first or senior authored papers in *Nature Genetics*, *Lancet Respiratory Medicine*, *Nature Communications, Genome Biology*, *Genome Research, NAR, AJRCCM, JACI, PNAS* as well as collaborative papers in *NEJM*, *Nature, Nature Genetics, Immunity and Nature Immunology.* He has served as a grant reviewer for NIH, French and Hong Kong research agencies. He is a reviewer for many top journals such as *NEJM*, *Nature Genetics, Nature Methods,* and *Nature Biotechnology.* He serves in the editorial board of *Genome Biology* and *American Journal of Respiratory Cell and Molecular Biology* and in the advisory board of American Thoracic Society. He has mentored over 20 pre- and post-doctoral fellows, who won multiple awards including two K01 grants from NIH under his supervision.

Eminent Scholar Talk
Wei Chen, PhD
Tuesday, August 10, 2021
8:40 – 9:00 PM (US EDT)

Title: Understanding Age-related Macular Degeneration in the Era of Big Data

Abstract: Age-related Macular Degeneration (AMD) is a multifactorial irreversible retina disease and the leading cause of blindness in the developed world. Multiple factors including aging, genetics, and smoking are associated with AMD development and its progression. Successful genome-wide association studies (GWAS) of AMD have identified over 30 genes that are significantly associated with advanced AMD including dry and wet subtypes. Supported by the National Eye Institute and other resources, my group has assembled several large-scale image and genetics datasets including tens of thousands of individuals with hundreds of thousands of color fundus images. The combination of wealthy genetics and fundus image data, plus the well-characterized clinical phenotypes provides unprecedented opportunities to explore novel directions for studying retinal disease. In this talk, I will present our recent computational work to address several key issues and challenges in the analysis of such multi-modal data. I will discuss several models to predict AMD risk and progression using genetics data, image data, or both. We show that statistical and deep learning approaches are critical in understanding AMD pathogenesis and predicting disease progression. Our methods and findings will have potential to enhance the early prevention and current clinical management of the disease and provide insights for novel precision treatment development.

A Live Tutorial on Molecular Evolutionary Genetics Analysis (MEGA) Tutorial (2.5 hours)

Sudhir Kumar[1], Ph.D and the laboratory members
[1]Professor and Director, Institute for Genomics and Evolutionary Medicine, Temple University

Abstract

The Molecular Evolutionary Genetics Analysis (MEGA) software contains an extensive repertoire of methods and tools for phylogenetic, phylogenomic, and phylomedicine analyses. MEGA's have grown with the addition of new methods, tools, and interfaces, resulting in a modern integrated software suite for comparative sequence analysis. It is now available in natively compiled applications with a rich Graphical User Interface (GUI) and Command-Line (CC) for Microsoft Windows, Linux, and macOS from www.megasoftware.net. In this workshop, we will provide a live tutorial demonstrating how to use MEGA for assembling sequence alignments, inferring phylogenetic trees, selecting substitution models, estimating genetic distances, inferring ancestral sequences, computing timetrees, and testing for natural selection. The workshop will begin with a presentation describing the brief history of MEGA, which will be followed by tutorials on (a) choosing, acquiring, and aligning sequences, (b) steps involved in inferring phylogenies from sequences sequence alignments or genetic distances, (c) the procedure of estimating divergence times, (d) the estimation of neutral evolutionary probabilities and tests of selection in MEGA. This will be followed by an introduction to the command-line version of MEGA for high-throughput and iterative data analysis. Presenters will answer questions during and after each tutorial segment.

Statistical Methods for Allele-Specific Expression Analysis Using RNA Sequencing Data Workshop (1.5 hours)

Jiaxin Fan[1,2], Ph.D, Rui Xiao[1], Ph.D
[1]Department of Biostatistics, Epidemiology and Informatics, The University of Pennsylvania Perelman School of Medicine
[2]The Food and Drug Administration

Abstract

Allele-specific gene expression (ASE) analysis, an alternative and complementary approach to eQTL analysis, is a powerful tool for identifying variation in gene expression. ASE quantifies the relative expression of two alleles in a diploid individual, and the imbalance of expression of the two alleles may explain phenotypic variation and disease pathophysiology. ASE is driven by cis-regulatory variants located near a gene. Since the two alleles used to measure ASE are expressed in the same cellular environment and genetic background, they can serve as internal controls and eliminate the influence of trans-acting genetic and environmental factors.

In the first session of this workshop (Dr. Xiao), we will focus on statistical methods for ASE analysis using RNA sequencing (RNA-seq) and single-cell RNA-seq (scRNA-seq) data, which provide allele-specific read counts distinguished by heterozygous sites. Specifically, we will first introduce a statistical model for detection of gene-level ASE across multiple individuals in a population under one clinical condition, as well as ASE difference between two clinical conditions. ASE patterns may vary across cell types. To better identify cellular targets of disease, we will next introduce a recently developed statistical method to characterize cell-type-specific ASE in bulk RNA-seq data by incorporating cell type composition information inferred from external scRNA-seq data. This method is extended to detect genes whose cell-type-specific ASE are associated with clinical factors by modeling covariate effect. Since having accurate cell type proportion estimate is critical for inferring the cell-type-specific ASE in bulk RNA-seq data, we will introduce MuSiC2, an iterative weighted non-negative least squares regression method, to deconvolve cell types in multi-condition bulk tissue RNA-seq data using scRNA-seq data from a single condition as reference.

In the second session of the workshop (Dr. Fan), we will offer the tutorials of how to use the statistical software packages for the three methods with data examples and annotation of the output.

With the growing popularity of RNA-seq and scRNA-seq, we believe these methods will provide a set of valuable tools for transcriptomics research. Results from the

analyses using these tools will offer insights on gene regulation and elucidate its relationship to human diseases.

Advanced computational analyses of single-cell RNA sequencing data
Tutorial (2.5 hours)

Yulin Dai[1], Ph.D, Hyun-Hwan Jeong[1], Ph.D
[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science at Houston

Abstract

Single-cell RNA-sequencing (scRNA-seq) technologies have rapidly advanced our understanding of transcriptome dynamics at an unprecedented resolution. With the wide usage of these technologies, numerous computational tools have been developed to address different biological questions. Due to the complexity of analysis, an improper combination of tools may produce an inaccurate result. Therefore, a standardized analysis pipeline for scRNA-seq analysis is strongly needed to produce a consistent and accurate result. In this tutorial, we will introduce how to pre-process the single-cell RNA-seq data, including quality control, normalization, data integration, clustering, and cell-type classification, as well as the cell- and gene-level downstream analysis, including differentially expressed gene analysis, trajectory inference, and cell-cell communication. We will provide the "best-practical recommendations" of the scRNA-seq analysis pipeline based on several frequent scenarios. Lastly, we will demonstrate how to integrate these steps into an R and Python workflow, giving the audience the convenience to replicates this analysis pipeline easily.

No Boundary thinking in bioinformatics
Workshop (1.5 hours)

Scott Williams[1], Ph.D, Weihua Guan[2], Ph.D., and other invited speakers
[1]Professor, Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University
[2]Associate Professor, Division of Biostatistics, School of Public Health, The University of Minnesota

Abstract

The rapid accumulation of large datasets in biomedicine can lead to substantial improvements in disease prediction, prevention and treatment. It can also allow us to better understand complex biological processes in all domains of the living world. As more data is being generated, collected, and distributed there is an increasing need to better formulate questions and problems that can help us improve our health and understanding of basic biological process. Traditional analyses have been based on explicit a priori hypotheses and statistical testing defined within pre-existing knowledge silos but with the explosion of data and computer power to assess these data there are new and exciting opportunities to change the frontiers of scientific research and industrial innovation and have new paradigms be defined not by classical approaches and ways of defining and designing analyses but by ones that are unbounded. We term this No-Boundary Thinking (NBT).

No-Boundary Thinking defines problems and solutions without of prior siloed thinking to address real and pressing scientific challenges. No-Boundary Thinking differs from multidisciplinary, interdisciplinary, or transdisciplinary research in that it posits that prior boundaries are in essence artificial constructs.

Our NBT workshop will promote No-Boundary Thinking in computational biology and bioinformatics. During the workshop, we will discuss the scientific challenges in different areas of computational biology and bioinformatics and use the discussion to refine the concept of NBT and how it can be used to re-assess complex biomedical problems.

Three invited speakers will cover the following topics:

- Speaker 1: Steven F. Jennings, University of Arkansas at Little Rock, "No Boundary Thinking: Solving Problems That Matter"
- Speaker 2: Philip R.O. Payne, Washington University in St. Louis, "Knowledge Integration"

- Speaker 3: Jonathan Stubblefield, Arkansas State University, "KITS19 Kidney Tumor Segmentation: Competition and Paper"

pHisPred: A Tool for the Identification of Histidine Phosphorylation Sites by Integrating Amino Acid Patterns and Properties

Jian Zhao[1], Minhui Zhuang[1], Jingjing Liu[1], Meng Zhang[1], Cong Zeng[1], Bin Jiang[2], Jing Wu[3*], Xiaofeng Song[1*]

[1] Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China.
[2] College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.
[3] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China.

[*] Corresponding authors

Background: Protein histidine phosphorylation (pHis) plays critical roles in prokaryotic signal transduction pathways and various eukaryotic cellular processes. It is estimated to account for 6%~10% of the phosphoproteome, however only hundreds of pHis sites have been discovered to date. Due to the inherent disadvantages of experimental methods, it is an urgent task for developing efficient computational approaches to identify pHis sites.

Results: Here, we present a novel tool, pHisPred, for accurately identifying pHis sites from protein sequences. We manually collected the largest number of experimental validated pHis sites to build benchmark datasets. Using randomized 10-fold CV, the weighted SVM-RBF model shows the best performance than other four commonly used classification models (LR, KNN, RF, and MLP). From ten thousands of features, 140 and 150 most informative features were individually selected out for eukaryotic and prokaryotic models. The average AUC and F1-score values of pHisPred were (0.81, 0.40) and (0.78, 0.46) for 10-fold CV on the eukaryotic and prokaryotic training datasets, respectively. In addition, pHisPred significantly outperforms other tools on testing datasets, in particular on the eukaryotic one.

Conclusion: We implemented a python program of pHisPred, which is freely available for non-commercial use at https://github.com/xiaofengsong/pHisPred. Moreover, users can use it to train new models with their own data.

Mining Functional Gene Modules by Multi-View NMF of Phenome-Genome Association

Xu Jin[1#], WenQian He[1#], MingMing Liu[1], Lin Wang[1], YaoGong Zhang[1], YingJie Xu[1], Ling Ma[1], YaLou Huang[2] and MaoQiang Xie[1*]

[1] College of Software, Nankai University, TianJin, China.
[2] TianJin International Joint Academy of Biomedicine, TianJin, China.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: Mining functional gene modules from genomic data is an important step to detect gene members of pathways or other relations such as protein-protein interactions. This work explores the plausibility of detecting functional gene modules by factorizing gene-phenotype association matrix from the phenotype ontology data rather than the conventionally used gene expression data. Recently, the hierarchical structure of phenotype ontologies has not been

sufficiently utilized in gene clustering while functionally related genes are consistently associated with phenotypes on the same path in phenotype ontologies.

Results: This work demonstrates a hierarchical Nonnegative Matrix Factorization (NMF) framework, called Consistent Multi-view Nonnegative Matrix Factorization (CMNMF), which factorizes genome-phenome association matrix at consecutive levels of the hierarchical structure in phenotype ontology to mine functional gene modules. CMNMF constrains the gene clusters from the association matrices at two consecutive levels to be consistent since the genes are

annotated with both the child-level phenotypes and the parent-level phenotypes in two levels. CMNMF also restricts the identified gene clusters to be densely connected in the phenotype ontology hierarchy. In the experiments on mining functionally related genes from mouse phenotype ontology and human phenotype ontology, CMNMF effectively improves clustering performance over the baseline methods. Gene ontology enrichment analysis is also conducted to verify its practical effectiveness to reveal meaningful gene modules.

Conclusions: Utilizing the information in the hierarchical structure of phenotype ontology, CMNMF can identify functional gene modules with more biological significance than conventional methods. CMNMF can also be a better tool for predicting members of gene pathways and protein-protein interactions.

Signaling interaction link prediction using deep graph neural networks integrating protein-protein interactions and omics data

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

[1] Institute for Informatics (I2), [2] Electrical and Systems Engineering Department, [3] Department of Pediatrics, Washington University School of Medicine, [4] Computer science, Washington University in St. Louis, St. Louis, MO, USA.

[*] Corresponding author

Uncovering signaling links or cascades among proteins that potentially regulate tumor development and drug response is one of the most critical and challenging tasks in cancer molecular biology. Inhibition of the targets on the core signaling cascades can be effective as novel cancer treatment regimens. However, signaling cascades inference remains an open problem, and there is a lack of effective computational models. The widely used gene co expression network (no-direct signaling cascades) and shortest-path based protein-protein interaction (PPI) network analysis (with too many interactions, and did not consider the sparsity of signaling cascades) were not specifically designed to predict the direct and sparse signaling cascades. To resolve the challenges, we proposed a novel deep learning model, deepSignalingLinkNet, to predict signaling cascades by integrating transcriptomics data and copy number data of a large set of cancer samples with the protein-protein interactions (PPIs) via a novel deep graph neural network model. Different from the existing models, the proposed deep learning model was trained using the curated KEGG signaling pathways to identify the informative omics and PPI topology features in the data-driven manner to predict the potential signaling cascades. The validation results indicated the feasibility of signaling cascade prediction using the proposed deep learning models. Moreover, the trained model can potentially predict the signaling cascades among the new proteins by transferring the learned patterns on the curated signaling pathways. The code was available at: https://github.com/fuhaililab/deepSignalingPathwayPrediction.

eSMC: a statistical model to infer admixture events from individual genomics data

Yonghui Wang[1,3#], Zicheng Zhao[2,3#], Xinyao Miao[2,4#], Yinan Wang[5], Xiaobo Qian[1*], Lingxi Chen[2], Changfa Wang[4] and Shuaicheng Li[2*]

[1] Liaocheng Research Institute of Donkey High-Efficiency Breeding and Ecological Feeding, Liaocheng University, 252059 Liaocheng, China.
[2] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Ave, Kowloon Tong, Hong Kong, China.
[3] The Byoryn Technology Co., Ltd, 518122 Shenzhen, China.
[4] School of forensic and medicine, Xi’an Jiaotong University, Xi’an, 710004 Xi’an, Shaanxi, P.R.China.
[5] The BGI Education Center, University of Chinese Academy of Sciences, 518083 Shenzhen, China.

# These authors contributed equally to this work
* Corresponding authors

Background: Inferring historical population admixture events yield essential insights in understanding a species demographic history. Methods are available to infer admixture events in demographic history with extant genetic data from multiple sources. Due to the deficiency in ancient population genetic data, there lacks a method for admixture inference from a single source. Pairwise Sequentially Markovian Coalescent (PSMC) estimates the historical effective
population size from lineage genomes of a single individual, based on the distribution of the most recent common ancestor between the diploid's alleles. However, PSMC does not infer the admixture event.

Results: Here, we proposed eSMC, an extended PSMC model for admixture inference from a single source. We evaluated our model's performance on both in silico data and real data. We simulated population admixture events at an admixture time range from 5 kya to 100 kya (5 years/generation) with population admix ratio at 1:1, 2:1, 3:1, and 4:1, respectively. The root means the square error is ±7.61 kya for all experiments. Then we implemented our method to infer the historical admixture events in human, donkey and goat populations. The estimated
admixture time for both Han and Tibetan individuals range from 60 kya to 80
kya (25 years/generation), while the estimated admixture time for the domesticated donkeys and the goats ranged from 40 kya to 60 kya (8 years/generation) and 40 kya to 100 kya (6 years/generation), respectively. The estimated admixture times were concordance to the time that
domestication occurred in human history.

Conclusion: Our eSMC effectively infers the time of the most recent admixture event in history from a single individual's genomics data. The source code of eSMC is hosted at https://github.com/zachary-zzc/eSMC.

Identifying genes associated with brain volumetric differences through tissue specific transcriptomic inference from GWAS summary data

Hung Mai[1,2], Jingxuan Bao[1,3], Paul M. Thompson[4], Dokyoon Kim[1], Li Shen[1*]

[1] Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
[2] School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA.
[3] School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA.
[4] Imaging Genetics Center, Stevens Institute for Neuroimaging & Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.

[*] Corresponding author

Background: Brain volume has been widely studied in the neuroimaging field, since it is an important and heritable trait associated with brain development, aging and various neurological and psychiatric disorders. Genome-wide association studies (GWAS) have successfully identified numerous associations between genetic variants such as single nucleotide polymorphisms (SNPs) and complex traits like brain volume. However, it is unclear how these genetic variations influence regional gene expression levels, which may subsequently lead to phenotypic changes. S-PrediXcan is a tissue-specific transcriptomic data analysis method that can be applied to bridge this gap. In this work, we perform an S-PrediXcan analysis on GWAS summary data from two large imaging genetics initiatives, the UK Biobank (UKB) and Enhancing Neuroimaging Genetics through Meta Analysis (ENIGMA), to identify tissue-specific transcriptomic effects on two closely related brain volume measures: total brain volume (TBV) and intracranial volume (ICV).

Results: As a result of the analysis, we identified 10 genes that are highly associated with both TBV and ICV. Nine out of 10 genes were found to be associated with TBV in another study using a different gene-based association analysis. Moreover, most of our discovered genes were also found to be correlated with multiple cognitive and behavioral traits. Further analyses revealed the protein-protein interactions, associated molecular pathways and biological functions that offer insight into how these genes function and interact with others.

Conclusions: These results confirm that S-PrediXcan can identify genes with tissue-specific transcriptomic effects on complex traits. The analysis also suggested novel genes whose expression levels are related to brain volumetric traits. This provides important insights into the genetic mechanisms of the human brain.

SCSilicon: a tool for synthetic single-cell DNA sequencing data generation

Xikang Feng[1*#] and Lingxi Chen[2#]

[1] School of Software, Northwestern Polytechnical University, Xi'an, 710072 Shaanxi, China.
[2] Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China.

[#] These authors contributed equally to this work
[*] Corresponding authors

Background: Single-cell DNA sequencing is getting indispensable in the study of cell-specific cancer genomics. The performance of computational tools that tackle single-cell genome aberrations may be nevertheless undervalued or overvalued, owing to the insufficient size of benchmarking data. In silicon simulation is a cost-effective approach to generate as many single-cell genomes as possible in a controlled manner to make reliable and valid benchmarking.

Results: This study proposes a new tool, SCSilicon, which efficiently generates single-cell in silicon DNA reads with minimum manual intervention. SCSilicon automatically creates a set of genomic aberrations, including SNP, SNV, Indel, and CNV. Besides, SCSilicon yields the ground truth of CNV segmentation breakpoints and subclone cell labels. We have manually inspected a series of synthetic variations. We conducted a sanity check of the start-of-art single-cell CNV callers and found SCYN was the most robust one.

Conclusions: SCSilicon is a user-friendly software package for users to develop and benchmark single-cell CNV callers. Source code of SCSilicon is available at https://github.com/xikanfeng2/SCSilicon.

McSNAC: A software to approximate first-order signaling networks from mass cytometry data

Darren Wethington[1,2#], Sayak Mukherjee[7], Jayajit Das[1-6#]

[1] Battelle Center for Mathematical Medicine, Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, Ohio.
[2] Biomedical Sciences Graduate Program, [3] Department of Pediatrics, [4] Pelotonia Institute for Immuno-Oncology, [5] Department of Biomedical Informatics, Wexner College of Medicine, [6] The Biophysics Graduate Program, The Ohio State University, Columbus, Ohio.
[7] Battelle Memorial Institute, Columbus, Ohio.

[*] Corresponding authors

Mass cytometry (CyTOF) gives unprecedented opportunity to simultaneously measure up to 40 proteins in single cells, with a theoretical potential to reach 100 proteins. This high-dimensional single-cell information can be very useful to dissecting mechanisms of cellular activity. In particular, measuring abundances of signaling proteins like phospho-proteins can provide detailed information on the dynamics of single-cell signaling processes. With a proper computational analysis, timestamped CyTOF data of signaling proteins could help develop predictive and mechanistic models for signaling kinetics. These models would be useful for predicting the effects of perturbations in cells, or comparing signaling networks across cell groups. We propose our Mass cytometry Signaling Network Analysis Code, or

McSNAC, a new software capable of reconstructing signaling networks and estimating their kinetic parameters from CyTOF data.

McSNAC approximates signaling networks as a network of first-order reactions between proteins. This assumption breaks down often as signaling reactions can involve binding and unbinding, enzymatic reactions, and other nonlinear constructions. Furthermore, McSNAC may be limited to approximating indirect interactions between protein species, as cytometry experiments are only able to assay a small fraction of the protein species that are involved in signaling. We carry out a series of in silico experiments here to show that 1) McSNAC is capable of accurately estimating the ground-truth model in a scalable manner when given data originating from a first-order system; 2) McSNAC is capable of qualitatively predicting outcomes to perturbations of species abundances in simple second-order reaction models and in a complex in silico nonlinear signaling network in which some proteins are unmeasured. These findings demonstrate that McSNAC can be a valuable screening tool for generating models of signaling networks from timestamped CyTOF data.

FSF-GA: A Feature Selection Framework for Phenotype Prediction Using Genetic Algorithms

Mohammad Erfan Mowlaei[1] and Xinghua Shi[1*]

[1] Department of Computer and Information Sciences, Temple University, 925 N. 12th Street, 19122 Philadelphia, PA, USA.

[*] Corresponding author

Background: Genome-Wide Association Studies (GWAS) involve scanning genetic markers across genomes in order to find associations of genetic variants and human phenotypes. Phenotype prediction is one of pivotal tasks in this field of study that helps scientists understand diseases or phenotypic differences introduced by genetic factors. There has been numerous research performed in this field. However, to this day, it is an open problem to understand the genetic contribution to complex phenotypes including common diseases due to the complexity between genotypes and phenotypes.

Results: In this paper, to perform quantitative phenotype prediction, we propose a novel feature selection framework for phenotype prediction utilizing a genetic algorithm (FSF-GA) that effectively reduces the feature space to identify genotypes in predicting a phenotype of interest. We provide a comprehensive vignette of our method and conduct extensive experiments using a widely used yeast dataset, which contains meticulously measured genotypes and respective quantitative traits.

Conclusions: According to the experimental results, our proposed FSF-GA method performs equally well compared with the well-known baseline methods, in terms of quantitative trait prediction, while it marks trait-associated loci. Experimental results show that our method delivers comparable phenotype prediction performance when benchmarked against baseline methods, while providing features selected for predicting the phenotype. These selected feature sets can be used to interpret genetic architecture contributing to phenotypic variation for the traits under investigation.

Deciphering the role of RNA structure in translation efficiency

Jianan Lin[1,2], Yang Chen[1], Yuping Zhang[3,4,5], Haifan Lin[6], and Zhengqing Ouyang[1*]

[1] Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA 01003, USA.
[2] The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA.
[3] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA.
[4] Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269, USA.
[5] Center for Quantitative Medicine, University of Connecticut, Farmington, CT 06030, USA.
[6] Yale Stem Cell Center, Department of Cell Biology, Yale University, New Haven, CT 06520, USA.

[*] Corresponding author

Background: RNA secondary structure has broad impact on the fate of RNA metabolism. The reduced stability of secondary structures in the translation-initiation region promotes the translation efficiency in both prokaryotic and eukaryotic species. However, the inaccuracy of in silico folding and the focus on the coding region limits our understanding of the global relationship between the whole mRNA RNA structure and translation efficiency. Leveraging high-throughput RNA structure probing data in the whole mRNAs, we aim to investigate the role of RNA structures in different regions in mRNA translation regulation.

Results: Here, we analyze the influence of hundreds of sequence and structure features in the whole mRNAs on the translation efficiency in the mouse embryonic stem cells (mESCs) and zebrafish developmental stages. Our findings suggest that overall in vivo RNA structure has a higher relative importance in predicting translation efficiency than in vitro RNA structure does in both mESCs and zebrafish cells. Furthermore, in vivo RNA structure in the 3’ UTR has the strongest influence on translation efficiency among all structure features in mESCs but not zebrafish cell. Instead, in vivo RNA structure in the 5’ UTR and coding regions around the translation initiation sites has weak but clearly higher influence on translation efficiency than in vitro structure does in zebrafish cells.

Conclusions: Our results suggest the openness of the 3' UTR promotes the translation efficiency in both mice and zebrafish, with the in vivo structure in 3' UTR is more important in mice than in zebrafish. This reveals a novel role of RNA secondary structure on translational regulation.

Integrative Analysis of Summary Data from GWAS and eQTL Studies Implicates Genes Differentially Expressed in Alzheimer's Disease

Brian Lee[1], Xiaohui Yao[1], Li Shen[1*] and for Alzheimer's Disease Neuroimaging Initiative[2]

[1] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[*] Corresponding author

Background: Although genome-wide association studies (GWAS) have successfully located various genetic variants susceptible to Alzheimer's Disease (AD), it is still unclear how specific variants interact with genes and tissues to elucidate pathologies associated with AD. Summary-data-based Mendelian Randomization (SMR) addresses this problem through an instrumental variable approach that integrates data from independent GWAS and expression quantitative trait locus (eQTL) studies in order to infer a causal effect of gene expression on a trait.

Results: Our study employed the SMR approach to integrate a set of meta-analytic cis-eQTL information from the Genotype-Tissue Expression (GTEx), CommonMind Consortium (CMC), and Religious Orders Study and Rush Memory and Aging Project (ROS/MAP) consortiums with three sets of meta-analysis AD GWAS results.

Conclusions: Our analysis identified twelve total gene probes (associated with twelve distinct genes) with a significant association with AD. Four of these genes survived a test of pleiotropy from linkage (the HEIDI test). Three of these genes -- RP11-385F7.1, PRSS36, and AC012146.7 -- have not yet been reported differentially expressed in the brain in the context of AD, and thus are the novel findings warranting further investigation.

CAISC: A Software to Integrate Copy Number Variation and Single Nucleotide Mutations for Genetic Heterogeneity Profiling and Subclone Detection by Single-cell RNA Sequencing

Jeerthi Kannan[1#], Liza Mathews[1#], Zhijie Wu[1], Neal S Young[1], Shouguo Gao[1*]

[1] Hematopoiesis and Bone Marrow Failure Laboratory, Hematology Branch, NHLBI, National Institutes of Health, Bethesda, Maryland 20892, USA.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: Although both copy number variations (CNVs) and single nucleotide variations (SNVs) detected by single cell RNA sequencing (scRNA-seq) are used to study intratumor heterogeneity and detect clonal groups, a software that integrates these two types of data in the same cells is unavailable.

Results: We developed Clonal Architecture with Integration of SNV and CNV (CAISC), an R package for scRNA-seq data analysis that clusters single cells into distinct subclones by integrating CNV and SNV genotype matrices using an entropy weighted approach. The performance of CAISC was tested on simulation data and four real datasets, which confirmed its high accuracy in sub-clonal identification and assignment, including subclones which cannot be identified using one type of data alone. Furthermore, integration of SNV and CNV allowed for accurate examination of expression change between subclones, as demonstrated by the results from trisomy 8 clones of the myelodysplastic syndromes (MDS) dataset.

Conclusions: CAISC is a powerful tool for integration of CNV and SNV data from scRNA-seq to identify clonal clusters with better accuracy than obtained from a single type of data. CAISC allows a user to interactively examine clonal assignments.

The Versatile Alignment Tool (VAT): A High-Performance Multi-Purpose Short Sequence Mapping Toolkit

Cuncong Zhong[1*], Xiangtao Liu[2]

[1] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA.

[2] Tianjia Genomes Tech. Co. LTD, Hefei, Anhui 238014, China.

[*] Corresponding author

Background: Biological sequence analysis plays a significant role in molecular biology research, especially with the increasing adoption of the low-cost, high-throughput next-generation sequencing (NGS) technology in the past decade. As data volume increases, traditional multi-purpose aligners (such as BLAST) are being replaced with more efficient special-purpose sequence mapping/alignment tools such as BWA and DIAMOND. However, efficient multi-purpose aligners are still in need to simplify comprehensive sequence analysis pipelines and to align sequences that are not generated from traditional sequencing protocols. To meet this demand, we develop an efficient multi-purpose aligner called the Versatile Alignment Tool (VAT); it currently supports both singular and chimeric short nucleotide and peptide sequence mapping.

Results: VAT contained three major algorithmic components: seeding, chaining, and gap-filling. VAT adopted a double-indexed suffix-array data structure for its seeding component and a dynamic programming-based chaining component, together the enables high efficiency and flexibility of VAT. Our benchmark results showed that VAT's performance (in terms of both alignment quality and efficiency) was comparable with its state-of-the-art special-purpose competitors. VAT also demonstrated unique advantages when n mapping chimeric reads.

Conclusion: In this work, we have developed a consolidated algorithmic framework for a variety of alignment modes. The resulted multi-purpose aligner VAT will have potential applications in integrative sequence analysis and handling novel NGS data types with previously unseen characteristics. VAT is implemented in GNU C++ and freely available from https://sourceforge.net/projects/vat-aligner.

Neural Representations of Cryo-EM Maps and a Graph-Based Interpretation

Nathan Ranno[1], Dong Si[1*]

[1] Department of Computing & Software Systems, University of Washington, Bothell, WA, USA.

[*] Corresponding author

Advances in imagery at atomic and near-atomic resolution, such as cryogenic electron microscopy (cryo-EM), have led to an influx of high resolution images of proteins and other macromolecular structures to data banks worldwide. Producing a protein structure from the discrete voxel grid data of cryo-EM maps involves interpolation into the continuous spatial domain. We present a novel data format called the neural cryo- EM map, which is formed from a set of neural networks that accurately parameterize cryo-EM maps and provide native, spatially continuous data for density and gradient. As a case study of this data format, we create graph-based interpretations of high resolution experimental cryo-EM maps. Normalized cryo-EM map values interpolated using the non-linear neural cryo-EM format are more accurate, consistently scoring less than 0.01 mean absolute error, than a conventional tri-linear interpolation, which scores up to 0.12 mean absolute error. Our graph-based interpretations of 115 experimental cryo-EM maps from 1.15 to 4.0 Å resolution provide high coverage of the underlying amino acid residue locations, while accuracy of nodes is correlated with resolution. The nodes of graphs created from atomic resolution maps (higher than 1.6 Å) provide greater than 99% residue coverage as well as 85% full atomic coverage with a mean of than 0.19 Å root mean squared deviation (RMSD). Other graphs have a mean 84% residue coverage with less specificity of the nodes due to experimental noise and differences of density context at lower resolutions. The fully continuous and differentiable nature of the neural cryo-EM map enables the adaptation of the voxel data to alternative data formats, such as a graph that characterizes the atomic locations of the underlying protein or macromolecular structure. Graphs created from atomic resolution maps are superior in finding atom locations and may serve as input to predictive residue classification and structure segmentation methods. This work may be generalized for transforming any 3D grid-based data format into non-linear, continuous, and differentiable format for the downstream geometric deep learning applications.

LENRM: a new noise reduction method based on local expansion for detecting overlapping protein complexes

Lei Xue[1], Xu Qing Tang[1*]

[1] School of Science, Jiangnan University, China.

[*] Corresponding author

Background: Overlapping protein complexes are increasingly important for us to understand biological systems and promote treatment of disease. In fact, there are two ways, the experimental method and the computational method, for detecting the overlapping protein complexes. Because the experimental method is difficult to verify the authenticity of the inferred complexes, the computational method for discovering the overlapping protein complexes becomes very important. Particularly, we can calculate the overlapping protein complexes from their protein-protein interaction (PPI) networks. Nevertheless, owing to the large amount of noise and unreliable estimates of interactions in the initial PPI network, to obtain the overlapping protein complexes through computational way is still a challenging problem.

Results: In this paper, a noise reduction method based on local expansion, LENRM, is developed to mine protein function modules. By removing inter-modules interactions and simulating undiscovered connections, our method eliminates high amount of false positive and false negative noise in PPI network respectively by modifying the resolution of the fitness function. Our approach is tested on several benchmark datasets and compared with nine known algorithms, which indicates the superior performance of it both in complexes' quantity and accuracy. LENRM performs best on 6 out of 8 datasets based on the ACC, MMR and Fraction.

Conclusion: In summary, we propose a new method by deleting inter-module interactions in large PPI networks, to extract protein complexes. The protein complexes discovered with LENRM method overall show significantly better agreement with the real complexes than nine existing methods, which provides some help for the research of the protein complex.

CrisprVi: a software for visualizing and analyzing CRISPR sequences of prokaryotes

Lei Sun[1,2,3,4*], Fu Yan[1,2], Gongming Wang[1,2], Jinbiao Wang[1,2], Yun Li[1,2,5], Jinlin Huang[3,6]

[1] School of Information Engineering, Yangzhou University, Yangzhou, P.R. China.
[2] School of Artificial Intelligence, Yangzhou University, Yangzhou, P.R. China.
[3] Jiangsu Key Laboratory of Zoonosis, Yangzhou, P.R. China.
[4] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, P.R. China.
[5] Jiangsu Province Engineering Research Center of Knowledge Management and Intelligent Service, Yangzhou, P.R. China.
[6] Jiangsu Co-Innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonosis, Yangzhou, P.R. China.

[*] Corresponding author

Background: Clustered regularly interspaced short palindromic repeats (CRISPR) and their spacers are important components of prokaryotic CRISPR-Cas systems. In order to analyze the CRISPR loci of multiple genomes more intuitively and comparatively, here we propose a visualization analysis tool named CrisprVi.

Results: CrisprVi is a Python package consisting of a graphic user interface (GUI) for visualization, a module for commands parsing and data transmission, local SQLite and BLAST databases for data storage and a functions layer for data processing. CrisprVi can not only visually present information of CRISPR repeats and spacers, such as their orders on the genome, IDs, start and end coordinates, but also provide interactive operation for users to display, label and align the CRISPR sequences, which help researchers to investigate the locations, orders and components of the CRISPR sequences in a global

view. In comparison to other CRISPR visualization tools such as CRISPRviz and CRISPRStudio, CrisprVi not only improves the interactivity and effects of visualization, but also provides basic statistics of the CRISPR sequences, and the consensus sequences of DRs/spacers across the input strains can be inspected from a clustering heatmap based on the BLAST results of the CRIPSR sequences hitting against the genomes.

Conclusions: CrisprVi is a convenient tool for visualizing and analyzing the CRISPR sequences and it would be helpful for users to inspect novel CRISPR-Cas systems of prokaryotes.

Deep Multiview Learning to Identify Imaging-driven Subtypes in Mild Cognitive Impairment

Yixue Feng[1*], Mansu Kim[2], Xiaohui Yao[2], Kefei Liu[2], Qi Long[2], Li Shen[2*], and for the Alzheimer's Disease Neuroimaging Initiative[3]

[1] School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, USA.
[2] Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.
[3] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[*] Corresponding authors

Background: In Alzheimer's Diseases (AD) research, multimodal imaging analysis can unveil complementary information from multiple imaging modalities and further our understanding of the disease. One application is to discover disease subtypes using unsupervised clustering. However, existing clustering methods are often applied to input features directly, and could suffer from the curse of dimensionality with high-dimensional multimodal data. The purpose of our study is to identify multimodal imaging-driven subtypes in Mild Cognitive Impairment (MCI) participants using a multiview learning framework based on deep generalized CCA (DGCCA), to learn shared latent representation with low dimensions from 3 neuroimaging modalities.

Results: DGCCA applies non-linear transformation to input views using neural networks and is able to learn embeddings with low dimensions that capture more variance than its linear counterpart, generalized CCA (GCCA). We designed experiments to compare DGCCA embeddings with single modality features and GCCA embeddings by generating

2 subtypes from each feature set using unsupervised clustering. In our validation studies, we found that amyloid PET imaging has the most discriminative features compared with structural MRI and FDG PET which DGCCA learns from but not GCCA. DGCCA subtypes show differential measures in 5 cognitive assessments, 6 brain volume measures, and conversion to AD patterns. In addition, DGCCA MCI subtypes confirmed AD genetic markers with strong signals that existing late MCI group did not identify.

Conclusion: Overall, DGCCA is able to learn effective low dimensional embeddings from multimodal data by learning non-linear projections. MCI subtypes generated from DGCCA embeddings are different from existing early and late MCI groups and show most similarity with those identified by amyloid PET features. In our validation studies, DGCCA subtypes show distinct patterns in cognitive measures, brain volumes, and are able to identify AD genetic markers. These findings indicate the promise of the imaging-driven subtypes and their power in revealing disease structures beyond early and late stage MCI.

B-assembler: a circular bacterial genome assembler

Fengyuan Huang[1,3], Li Xiao[2], Min Gao[1,2], Ethan J Vallely[1], Kevin Dybvig[3,4], T. Prescott Atkinson[4], Ken B. Waites[5], Zechen Chong[1,3*]

[1] Informatics Institute, the University of Alabama at Birmingham, Birmingham, Alabama, 35294, United States of America.
[2] Department of Medicine, the University of Alabama at Birmingham, Birmingham, Alabama, 35294, United States of America.
[3] Department of Genetics, the University of Alabama at Birmingham, Birmingham, Alabama, 35294, United States of America.
[4] Department of Pediatrics, the University of Alabama at Birmingham, Birmingham, Alabama, 35233, United States of America.
[5] Department of Pathology, the University of Alabama at Birmingham, Birmingham, Alabama, 35233, United States of America.

[*] Corresponding author

Background: Accurate bacteria genome de novo assembly is fundamental to understand the evolution and pathogenesis of new bacteria species. The advent and popularity of Third-Generation Sequencing (TGS) enables assembly of bacteria genomes at an unprecedented speed. However, most current TGS assemblers were specifically designed for human or other species that do not have a circular genome. Besides, the repetitive DNA fragments in many bacterial genomes plus the high error rate of long sequencing data make it still very challenging to accurately assemble their genomes even with a relatively small genome size. Therefore, there is an urgent need for the development of an optimized method to address these issues.

Results: We developed B-assembler, which is capable of assembling bacterial genomes when there are only long reads or a combination of short and long reads. B-assembler takes advantage of the structural resolving power of long reads and the accuracy of short reads if applicable. It first selects and corrects the ultra-long reads to get an initial contig. Then, it collects the reads overlapping with the ends of the initial contig. This two-round assembling procedure along with optimized error correction enables a high- confidence and circularized genome assembly. Benchmarked on both synthetic and real sequencing data of several species of bacterium, the results show that both long- read-only and hybrid-read modes can accurately assemble circular bacterial genomes free of structural errors and have fewer small errors compared to other assemblers.

Conclusions: B-assembler provides a better solution to bacterial genome assembly, which will facilitate downstream bacterial genome analysis.

DISTEMA: distance map-based estimation of single protein model accuracy with attentive 2D convolutional neural network

Xiao Chen[1], Jianling Cheng[1*]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA.

[*] Corresponding author

Background: Estimation of the accuracy (quality) of protein structural models is important for both prediction and use of protein structural models. Deep learning methods have been used to integrate protein structure features to predict the quality of protein models. Inter-residue distances are key information for predicting protein's tertiary structures and therefore have good potentials to predict the quality of protein structural models. However, few methods have been developed to fully take advantage of predicted inter-residue distance maps to estimate the accuracy of a single protein structural model.

Result: We developed an attentive 2D convolutional neural network (CNN) with channel-wise attention to take only a raw difference map between the inter-residue distance map calculated from a single protein model and the distance map predicted from the protein sequence as input to predict the quality of the model. The network comprises multiple convolutional layers, batch normalization layers, dense layers, and Squeeze-and-Excitation blocks with attention to automatically extract features relevant to protein model quality from the raw input without using any expert-curated features. We evaluated DISTEMA's capability of selecting the best models for CASP13 targets in terms of ranking loss of GDT-TS score. The ranking loss of DISTEMA is 0.079, lower than several state-of-the-art

single-model quality assessment methods. The work demonstrates that using raw inter-residue distance information alone with deep learning can predict the quality of protein structural models reasonably well.

BVMHC: Bilateral and Variable Long Short Term Memory Networks Based Major Histocompatibility Complex Binding Prediction

Limin Jiang[1], Hui Yu[1], Jijun Tang[2,3], Fei Guo[2*], Yan Guo[1*]

[1] Comprehensive cancer center, Department of Internal Medicine, University of New Mexico, Albuquerque, NM, USA.
[2] School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China.
[3] Department of Computer Science, University of South Carolina, SC, 29208, USA.

[*] Corresponding authors

Background: As an important part of immune surveillance, Major Histocompatibility Complex (MHC) is a set of proteins that recognize foreign molecules. Computational prediction methods for MHC binding peptides have been developed. However, existing methods share the limitation of fixed peptide sequence length, which necessitates the training of models by peptide length or prediction with a length reduction technique.

Results: Using Bilateral and Variable Long Short-Term Memory neural network, we constructed BVMHC, an MHC class I and II binding prediction tool that is independent of peptide length. The performance of BVMHC was compared to seven MHC class I prediction tools and three MHC class II prediction tools using eight criteria in an independent validation dataset. BVMHC achieved the best performance in six of the eight criteria for MHC class I, and the best performance in all eight criteria for MHC class II, including accuracy and AUC. Furthermore, models for non-human species were also trained using the same strategy and made available for applications in mouse, chimpanzee, macaque, and rat.

Conclusion: BVMHC is a peptide length independent MHC class I and II binding predictor. Models from this study have been implemented in an online web portal for easy access and usage.

Integrative analysis of eQTL and GWAS summary statistics reveals transcriptomic alteration in Alzheimer brains

Pradeep Varathan[1], Priyanka Gorijala[1], Tanner Jacobson[2], Danai Chasioti[1], Kwangsik Nho[2], Shannon L Risacher[2], Andrew J Saykin[2] and Jingwen Yan[1,2*]

[1] Department of BioHealth Informatics, Indiana University Purdue University Indianapolis, Indianapolis, Indiana, USA.
[2] Department of Radiology and Imaging Sciences, School of Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA.

[*] Corresponding author

Large-scale genome-wide association studies have successfully identified many genetic variants significantly associated with Alzheimer's disease (AD), such as rs429358, rs11038106, rs723804, rs13591776, and more. The next key step is to understand the function of these SNPs and the downstream biology through which they exert the effect on the development of AD. However, this remains a challenging task due to the tissue-specific nature of transcriptomic and proteomic data and the limited availability of brain tissue. In this paper, instead of using coupled transcriptomic data, we performed an integrative analysis of existing GWAS findings and expression quantitative trait loci (eQTL) results from AD-related brain regions to estimate the transcriptomic alterations in AD brain. We used summary-based mendelian randomization method along with heterogeneity in dependent instruments method (HEIDI) and were able to identify 32 genes with potential altered levels in temporal cortex region. Among these, 10 of them were further validated using real gene expression data collected from temporal cortex region, and 19 SNPs from NECTIN and TOMM40 genes were found associated with multiple temporal cortex imaging phenotype.

AEDNav: Indoor navigation for locating Automated External Defibrillator

Gaurav Rao[1*], Vijay Mago[2], Pawan Lingras[1] and David W. Savage[3]

[1] Department of Mathematics & Computing Science, Saint Mary's University, Halifax, CA.
[2] Department of Computer Science, Lakehead University, Thunder Bay, CA.
[3] Northern Ontario School of Medicine, Thunder Bay, CA.

[*] Corresponding author

Background: In a sudden cardiac arrest, starting CPR and applying an AED immediately are the two highest resuscitation priorities. Many existing mobile applications have been developed to assist users in locating a nearby AED. However, these applications do not provide indoor navigation to the AED location. The time required to locate an AED inside a building due to a lack of indoor navigation systems will reduce the patient's chance of survival. The existing indoor navigation solutions either require special hardware, a large dataset or a significant amount of initial work. These requirements make these systems not viable for implementation on a large-scale.

Methods: The proposed system collects Wi-Fi information from the existing devices and the path's magnetic information using a smartphone to guide the user from a starting point to an AED. The information collected is processed using four techniques: turn detection method, Magnetic data pattern matching method, Wi-Fi fingerprinting method and Closest Wi-Fi location method to estimate user location. The user location estimations from all four techniques are further processed to determine the user's location on the path, which is then used to guide the user to the AED location. Results: The four techniques used in the proposed system Turn detection, Magnetic data pattern matching, Closest Wi-Fi location and Wi-Fi fingerprinting can individually achieve the accuracy of 80% with the error distance ± 9.4 meters, ± 2.4 meters, ± 4.6 meters , and ± 4.6 meters respectively. These four techniques, applied individually, may not always provide stable results. Combining these techniques results in a robust system with an overall accuracy of 80% with an error distance of ± 2.74 meters. In comparison, the proposed system's accuracy is higher than the existing systems that use Wi-Fi and magnetic data.

Conclusion: This research proposes a novel approach that requires no special hardware, large scale data or significant initial work to provide indoor navigation. The proposed system AEDNav can achieve an accuracy similar to the existing indoor navigation systems. Implementing this indoor navigation system could reduce the time to locate an AED and ultimately increase patient survival during sudden cardiac arrest.

Estimating the optimal linear combination of predictors using spherically constrained optimization

Priyam Das[1*], Debsurya De[2], Raju Maiti[3], Mona Kamal[4], Katherine A. Hutcheson[5], Clifton D. Fuller[4], Bibhas Chakraborty[3,6,7] and Christine B. Peterson[8]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
[2] Indian Statistical Institute, Kolkata, India.
[3] Centre for Quantitative Medicine, Duke-National University of Singapore Medical School, Singapore.
[4] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
[5] Department of Head and Neck Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
[6] Department of Statistics and Applied Probability, National University of Singapore, Singapore.
[7] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA.
[8] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[*] Corresponding author

Background: In the context of a binary classification problem, the optimal linear combination of continuous predictors can be estimated by maximizing the area under the receiver operating characteristic curve. For ordinal responses, the optimal predictor combination can similarly be obtained by maximization of the hypervolume under the manifold (HUM). Since the empirical HUM is discontinuous, non-differentiable, and possibly multi-modal, solving this maximization problem requires a global optimization technique. Estimation of the optimal coefficient vector using existing global optimization techniques is computationally expensive, becoming prohibitive as the number of predictors and the number of outcome categories increases.

Results: We propose an efficient derivative-free black-box optimization technique based on pattern search to solve this problem, which we refer to as Spherically Constrained Optimization Routine (SCOR). Through extensive simulation studies, we demonstrate that the proposed method achieves better performance than existing methods including the step-down algorithm. Finally, we illustrate the proposed method to predict the severity of swallowing difficulty after radiation therapy for oropharyngeal cancer based on radiation dose to various structures in the head and neck.

Conclusions: Our proposed method addresses an important challenge in combining multiple biomarkers to predict an ordinal outcome. This problem is particularly relevant to medical research, where it may be of interest to diagnose a disease with various stages of progression or a toxicity with multiple grades of severity. We provide the implementation of our proposed SCOR method as an R package, available online at https://CRAN.R-project.org/package=SCOR.

DENSEN: a convolutional neural network for estimating chronological ages from panoramic radiographs

Xuedong Wang[1, 2#], Xinyao Miao[1, 3, 4#], Yanle Liu[1#], Yin Chen[2#], Xiao Cao[1], Yuchen Zhang[1], Shuaicheng Li[4*] and Qin Zhou[1*]

[1] The Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases & Department of Implant Dentistry, College of Stomatology, Xi'an Jiaotong University., 710004 Xi'an, Shaanxi, P.R.China.
[2] The Byoryn Technology Co., Ltd, 518122 Shenzhen, China.
[3] School of forensic and medicine, Xi'an Jiaotong University, Xi'an, 710004 Xi'an, Shaanxi, P.R.China.
[4] City University of Hong Kong, 83 Tat Chee Ave, Kowloon Tong, Hong Kong, China.

[#] These authors contributed equally to this work
[*] Corresponding authors

Background: Age estimation from panoramic radiographs is a fundamental task in forensic sciences. Previous age assessment studies mainly focused on juvenile rather than elderly populations (>25 years old). Most proposed studies were statistical or scoring-based, requiring wet-lab experiments and professional skills, and suffering from low reliability.

Result: Based on Soft Stagewise Regression Network (SSR-Net), we developed DENSEN to estimate the chronological age for both juvenile and older adults, based on their orthopantomograms (OPTs, also known as orthopantomographs, pantomograms, or panoramic radiographs). We collected 1,903 clinical panoramic radiographs of individuals between three-year and eighty-five years old to train and validate the model. We evaluated the model by the mean absolute error (MAE) between the estimated age and ground truth. For different age groups, 3-11(children), 12-18(teens), 19-25(young adults), and 25+(adults), DENSEN produced MAEs as 0.6885, 0.7615, 1.3502, and 2.8770, respectively. Our results imply that the model works in situations where genders are unknown. Moreover, DENSEN has lower errors for the adult group (>25 years) than other methods. The proposed model is memory compact, consuming about 1.0 MB of memory overhead.

Conclusions: We presented a novel deep learning approach DENSEN to estimate a subject's age from a panoramic radiograph for the first time. Our approach required less laboratory work compared with existing methods. The package we developed is an open-source tool and applies to all different age groups.

Identification of Multimodal Brain Imaging Association via A Parameter Decomposition based Sparse Multi-view Canonical Correlation Analysis Method

Jin Zhang[1], Huiai Wang[1], Ying Zhao[1], Lei Guo[1], Lei Du[1*] and the Alzheimer’s Disease Neuroimaging Initiative

[1] School of Automation, Northwestern Polytechnical University, Xi’an, China.

[*] Corresponding author

Background: With the development of noninvasive imaging technology, collecting different imaging measurements of the same brain becomes more and more easily. These multimodal imaging data carry complementary information of the same brain, with both specific and shared information being intertwined. Within these multimodal data, it is essential to discriminate specific information from shared information since it is of benefit to comprehensively characterize brain diseases. While most existing methods are unqualified, in this paper, we propose a parameter decomposition based sparse multi-view canonical correlation analysis (DSMCCA) method. DSMCCA could identify both modality-shared and -specific information of multimodal data, leading to an in-depth understanding of complex pathology of brain disease.

Results: Compared with the SMCCA method, our method obtains higher correlation coefficients and better canonical weights on both synthetic data and real neuroimaging data. This indicates that, coupled with modality-shared and -specific feature selection, DSMCCA improves the multi-view association identification and shows meaningful feature selection capability with desirable interpretation.

Conclusions: The novel DSMCCA confirms that the parameter decomposition is a suitable strategy to identify both modality-shared and -specific imaging features. The multimodal association and the diverse information of multimodal imaging data enable us to better understand the brain disease such as Alzheimer’s disease.

An Integrated Interactive COVID-19 Dashboard for Individual Risk Analysis and Real-time Trend Analysis

Josh Voytek[1], Maria Maltepes[1], Anna Lengner[1], Jay S. Patel[1*], Huanmei Wu[1*]

[1] Department of Health Services Administration and Services, The College of Public Health, Temple University, Philadelphia, PA 19122, USA.

[*] Corresponding authors

Background: Although many dashboards have been developed for the COVID-19 pandemic, they have many limitations, ranging from a poor user experience to difficulty comparing different trends. Currently, no existing dashboard combines extensive data visualization on a national level and tailored risk predictions on the individual level. In this study, we developed an interactive dashboard to visualize COVID-19 data in the US in a way that is accessible, reliable, and easily understood while also featuring an infection and mortality risk calculator for individual users to try out. Methods We developed a dashboard that utilizes CDC datasets to generate graphs that visualize cases, deaths, RT-PCR testing, and vaccination efforts. The user also can overlay the trends of different states and territories, view the trends in any time frame since the data was first recorded (in weeks), and apply trendlines. We also developed a risk assessment tool that provides risk predictions for COVID-19 infection and mortality based on variables including gender, age, race/ethnicity, county and state location, behavior, and pre-existing medical conditions retrieved from the CDC, the NYTimes, and existing literature.

Results: The risk calculator outputs individualized infection and mortality rates for each of the user's inputs. Along with their risks, a summary is attached to explain the data sources and why there are apparent elevated risks for specific demographics. The graphs provide state-level data for the user, allowing them to manipulate the displayed data via several sorting mechanisms, and provide different forms of trendlines with their associated equations. The state data can also be compared side-by-side with a comparative graphs.

Conclusions: Our dashboard informs the user of current and past national COVID-19 trends in the US while also supplementing the users with the tools and information needed to discover significant correlations and create predictive models. A risk 51 assessment tool has also been integrated into the dashboard to provide users with individualized risk predictions based on their attributes.

Mining Comorbidities of Opioid Use Disorder from FDA Adverse Event Reporting System and Patient Electronic Health Records

Yiheng Pan[1], Rong Xu[1*]

[1] Case Western Reserve University, Cleveland, Ohio, USA.

[*] Corresponding author

Background: Opioid use disorder (OUD) has become an urgent health problem. People with OUD often experience comorbid medical conditions. Systematical approaches to identifying co-occurring conditions of OUD can facilitate a deeper understanding of OUD mechanisms and drug discovery. This study presents an integrated approach combining data mining, network construction and ranking, and hypothesis-driven case-control studies using patient electronic health records (EHRs).

Methods: First, we mined comorbidities from the US Food and Drug Administration Adverse Event Reporting System (FAERS) of 12 million unique case reports using frequent pattern-growth algorithm. The performance of OUD comorbidity mining was measured by precision and recall using manually curated known OUD comorbidities. We then constructed a disease comorbidity network using mined association rules and further prioritized OUD comorbidities. Last, novel OUD comorbidities were independently tested using EHRs of 75 million unique patients.

Results: The OUD comorbidities from association rules mining achieves a precision of 38.7% and a recall of 78.2 Based on the mined rules, the global DCN was constructed with 1916 nodes and 32,175 edges. The network-based OUD ranking result shows that 43 of 55 known OUD comorbidities were in the first decile with a precision of 78.2%. Hypothyroidism and type 2 diabetes were two top-ranked novel OUD comorbidities identified by data mining and network ranking algorithms. Based on EHR-based case-control studies, we showed that patients with OUD had significantly increased risk for hyperthyroidism (AOR = 1.46, 95% CI: 1.43-1.49, p-value < 0.001), hypothyroidism (AOR = 1.45, 95% CI: 1.42-1.48, p-value < 0.001), type 2-diabetes (AOR = 1.28, 95% CI: 1.26-1.29, p-value < 0.001), compared with individuals without OUD.

PheNominal: An EHR-Integrated Web Application for Structured Deep Phenotyping at the Point of Care

James M. Havrilla[1], Anbumalar Singaravelu[2], Dennis M. Driscoll[2], Leonard Minkovsky[2], Ingo Helbig[3-6], Livija Medne[7], Kai Wang[1,5,8], Ian Krantz[7], Bimal R. Desai[9*]

[1] Center for Cellular and Molecular Therapeutics, Children’s Hospital of Philadelphia, Philadelphia, PA, 19104, USA.
[2] Emerging Technology and Transformation Team, Information Services, Children’s Hospital of Philadelphia, Philadelphia, PA, 19104, USA.
[3] Division of Neurology, Children’s Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[4] The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, USA.
[5] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104 USA.
[6] Department of Neurology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, 19104 USA.
[7] Roberts Individualized Medical Genetics Center, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA.
[8] Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, 19104, USA.
[9] Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, 19104, USA.

[*] Corresponding author

Objective: Clinical phenotype information greatly facilitates genetic diagnostic interpretations pipelines in disease. While post-hoc extraction using natural language processing (NLP) on unstructured clinical notes continues to improve, there is a need to improve point-of-care collection of patient phenotypes. Therefore, we developed "PheNominal", a point-of-care web application, embedded within Epic electronic health record (EHR) workflows, to permit capture of standardized phenotype data.

Methods: Using bi-directional web services available within commercial EHRs, we developed a lightweight web application that allows users to rapidly browse and identify relevant terms from the Human Phenotype Ontology (HPO). Selected terms are saved discretely within the patient's EHR, permitting reuse both in clinical notes as well as in downstream diagnostic and research pipelines.

Results: In the 16 months since implementation, PheNominal was used to capture discrete phenotype data for over 1,500 individuals and 11,000 HPO terms during clinic and inpatient encounters for a genetic diagnostic consultation service within a quaternary-care pediatric academic medical center. An average of 7 HPO terms were captured per patient. Compared to a manual workflow, the average time to enter terms for a patient was reduced from 15 minutes to 5 minutes per patient, and there were fewer annotation errors.

Discussion: Modern EHRs support integration of external applications using application programming interfaces. We describe a practical application of these interfaces to facilitate deep phenotype capture in a discrete, structured format within a busy clinical workflow. Future versions will include a vendor-agnostic implementation using FHIR.

Conclusion: We describe pilot efforts to integrate structured phenotyping through controlled dictionaries into diagnostic and research pipelines, reducing manual effort for phenotype documentation and reducing errors in data entry.

Natural language processing to identify lupus nephritis phenotype in electronic health records

Yu Deng[1], Jennifer A. Pacheco[2], Anh Chung[1,6], Chengsheng Mao[1], Joshua C. Smith[3], Juan Zhao[3], Wei-Qi Wei[3], April Barnado[4], Chunhua Weng[5], Cong Liu[5], Adam Cordon[2], Jingzhi Yu[1], Yacob Tedla[1], Abel Kho[1], Rosalind Ramsey- Goldman[6], Theresa Walunas[1*#], Yuan Luo[1*#]

[1] Center for Health Information Partnerships, Feinberg School of Medicine, Northwestern University, Chicago, USA.
[2] Center for Genetic Medicine, Feinberg School of Medicine, Northwestern University, Chicago, USA.
[3] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, USA.
[4] Department of Medicine, Vanderbilt University Medical Center, Nashville, USA.
[5] Department of Biomedical Informatics, Columbia University, New York, USA.
[6] Department of Medicine/Rheumatology, Feinberg School of Medicine, Northwestern University, Chicago, USA.

[#] These authors contributed equally to this work
[*] Corresponding authors

Systemic lupus erythematosus (SLE) is a rare autoimmune disorder characterized by an unpredictable course of flares and remission with diverse manifestations. Lupus nephritis, one of the major disease manifestations of SLE for organ damage and mortality, is a key component of lupus classification criteria. Accurately identifying lupus nephritis in electronic health records (EHRs) would therefore benefit large cohort observational studies and clinical trials where characterization of the patient population is critical for recruitment, study design, and analysis. Lupus nephritis can be recognized through procedure codes and structured data, such as laboratory tests. However, other critical information documenting lupus nephritis, such as histologic reports from kidney biopsies and prior medical history narratives, require sophisticated text processing to mine information from pathology reports and clinical notes. In this study, we developed algorithms to identify lupus nephritis with and without natural language processing (NLP) using EHR data from the Northwestern Medicine Enterprise Data Warehouse (NMEDW). We developed four algorithms: a rule-based algorithm using only structured data (baseline algorithm) and three algorithms using different NLP models. The three NLP models are based on regularized logistic regression and use different sets of features including positive mention of concept unique identifiers (CUIs), number of appearances of CUIs, and a mixture of three components (i.e. a curated list of CUIs, regular expression concepts, structured data) respectively. The baseline algorithm and the best performed NLP algorithm were external

validated on a dataset from Vanderbilt University Medical Center (VUMC). Our best performing NLP model incorporating features from both structured data, regular expression concepts, and mapped concept unique identifiers (CUIs) improved F measure in both the NMEDW (0.41 vs 0.79) and VUMC (0.62 vs 0.96) datasets compared to the baseline lupus nephritis algorithm.

Expediting knowledge acquisition by a web framework for Knowledge Graph Exploration and Visualization (KGEV): a case study on COVID-19

Jacqueline Peng[1], David Xu[2], Ryan Lee[2], Siwei Xu[3], Yunyun Zhou[1*], Kai Wang[1,4*]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
[2] School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA.
[3] College of Arts & Sciences, Emory University, Atlanta, GA 30322, USA.
[4] Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.

[*] Corresponding authors

Background: Knowledges graphs (KGs) serve as a convenient framework for structuring knowledge. A number of computational methods have been developed to generate KGs from biomedical literature and use them for downstream tasks such as link prediction and questioning and answering. However, there is a lack of computational tools or web frameworks to support the exploration and visualization of the KG themselves, which would facilitate interactive knowledge discovery and formulation of novel biological hypotheses.
Method: We developed a web framework for Knowledge Graph Exploration and Visualization (KGEV), to construct and visualize KGs in five stages: triple extraction, triple filtration, metadata preparation, knowledge integration, and graph database preparation. The application has convenient user interface tools, such as node and edge search and filtering, data source filtering, neighborhood retrieval, shortest path calculation, that are performed by querying a backend graph database. Unlike other KGs, our framework allows fast retrieval of relevant texts supporting the relationships in the KG, thus allowing human reviewers to judge the reliability of the knowledge extracted.

Results: We demonstrated a case study of using the KGEV framework to perform research on COVID-19. The COVID-19 pandemic resulted in an explosion of relevant literature, making it challenging to make full use of the vast and heterogenous sources of information. We generated a COVID-19 KG with heterogenous information, including literature information from the CORD-19 dataset, as well as other existing knowledge from eight data sources. We showed the utility of KGEV in three intuitive case studies to explore and

query knowledge on COVID-19. A demo of this web application can be accessed at http://covid19nlp.wglab.org. Finally, we also demonstrated a turn-key adaption of the KGEV framework to study clinical phenotypic presentation of human diseases, illustrating the versatility of the framework.

Conclusion: In an era of literature explosion, the KGEV framework can be applied to many emerging diseases to support structured navigation of the vast amount of newly published biomedical literature and other existing biological knowledge in various databases. It can be also used as a general-purpose tool to explore and query gene-phenotype-disease-drug relationships interactively.

Disparities in Social Determinants among Performances of Mortality Prediction with Machine Learning for Sepsis Patients

Hanyin Wang[1], Yikuan Li[1], Andrew Naidech[2], Yuan Luo[1*]

[1] Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA.
[2] Department of Neurology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA.

[*] Corresponding author

Background: Sepsis is one of the most life-threatening circumstances for critically ill patients in the United States, while diagnosis of sepsis is challenging as a standardized criteria for sepsis identification is still under development. Disparities in social determinants of sepsis patients can interfere with the risk prediction performances using machine learning. Methods We analyzed a cohort of critical care patients from the Medical Information Mart for Intensive Care (MIMIC)-III database. Disparities in social determinants, including race, gender, marital status, insurance types and languages, among patients identified by six available sepsis criteria were revealed by forest plots with 95% confidence intervals. Sepsis patients were then identified by the Sepsis-3 criteria. Sixteen machine learning classifiers were trained to predict in-hospital mortality for sepsis patients on a training set constructed by random selection. The performance was measured by area under the receiver operating characteristic curve (AUC). The performance of the trained model was tested on the entire randomly conducted test set and each sub-population built based on each of the following social determinants: race, gender, marital status, insurance type, and language. The fluctuations in performances were further examined by permutation tests.

Results: We analyzed a total of 11,791 critical care patients from the MIMIC-III database. Within the population identified by each sepsis identification method, significant

differences were observed among sub-populations regarding race, marital status, insurance type, and language. On the 5,783 sepsis patients identified by the Sepsis-3 criteria statistically significant performance decreases for mortality prediction were observed when applying the trained machine learning model on Asian and Hispanic patients, as well as the Spanish-speaking patients. With pairwise comparison, we detected performance discrepancies in mortality prediction between Asian and White patients, Asians and patients of other races, as well as English-speaking and Spanish-speaking patients.

Conclusions: Disparities in proportions of patients identified by various sepsis criteria were detected among the different social determinant groups. The performances of mortality prediction for sepsis patients can be compromised when applying a universally trained model for each subpopulation. To achieve accurate diagnosis, a versatile diagnostic system for sepsis is needed to overcome the social determinant disparities of patients.

On the Role of Deep Learning Model Complexity in Adversarial Robustness for Medical Images

David Rodriguez[1], Tapsya Nayak[1], Yidong Chen[2,3], Ram Krishnan[1*] and Yufei Huang[1*]

[1] Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX, USA.
[2] Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX, USA.
[3] Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX, USA.

[*] Corresponding authors

Background: Deep learning (DL) models are highly vulnerable to adversarial attacks for medical image classification. An adversary could modify the input data in imperceptible ways such that a model could be tricked to predict, say, an image that actually exhibits malignant tumor to a prediction that it is benign. However, adversarial robustness of DL models for medical images is not adequately studied. DL in medicine is inundated with models of various complexity --- particularly, very large models. In this work, we investigate the role of model complexity in adversarial settings.

Results: Consider a set of DL models that exhibit similar performances for a given task. These models are trained in the usual manner but are not trained to defend against adversarial attacks. We demonstrate that, among those models, simpler models of reduced complexity show a greater level of robustness against adversarial attacks than larger models that often tend to be used in medical applications. On the other hand, we also show that once those models undergo adversarial training, the adversarial trained medical image

DL models exhibit a greater degree of robustness than the standard trained models for all model complexities.

Conclusion: The above result has a significant practical relevance. When medical practitioners lack the expertise or resources to defend against adversarial attacks, we recommend that they select the smallest of the models that exhibit adequate performance. Such a model would be naturally more robust to adversarial attacks than the larger models.

Application of Unsupervised Deep Learning Algorithms for Identification of Specific Clusters of Chronic Cough Patients from EMR Data

Wei Shao[1], Xiao Luo[2*], Zuoyi Zhang[1], Zhi Han[1,4], Vasu Chandrasekaran[3], Vladimir Turzhitsky[3], Vishal Bali[3], Anna R. Roberts[4], Megan Metzger[4], Jarod Baker[4], Carmen La Rosa[3], Jessica Weaver[3], Paul Dexter[1,4,5], and Kun Huang[1,4*]

[1] Indiana University School of Medicine, Indianapolis, USA.
[2] Purdue School of Engineering and Technology, IUPUI, Indianapolis, USA.
[3] Center for Observational and Real-World Evidence, Merck & Co., Inc., Kenilworth, NJ, USA.
[4] Regenstrief Institute, Inc., Indianapolis, IN, USA.
[5] Eskenazi Health, Indianapolis, IN, USA.

[*] Corresponding authors

Objective: Chronic cough (CC) affects approximately 10% of adults. The lack of ICD codes for chronic cough makes it challenging to apply supervised learning methods to predict the characteristics of chronic cough patients, thereby requiring the identification of chronic cough patients by other mechanisms.

Materials and Method: We developed a deep clustering algorithm with auto-encoder embedding (DCAE) to identify clusters of chronic cough patients based on data from a large cohort of 264,146 patients from the Electronic Medical Records (EMR) system. We constructed features using the diagnosis within the EMR, then built a clustering-oriented loss function directly on embedded features of the deep autoencoder to jointly perform feature refinement and cluster assignment. Lastly, we performed statistical analysis on the identified clusters to characterize the chronic cough patients compared to the non-chronic cough patients.

Results: The experimental results show that the DCAE model generated three chronic cough clusters and one non-chronic cough patient cluster. We found various diagnoses, medications, and lab tests highly associated with chronic cough patients by comparing the chronic cough cluster with the nonchronic cough cluster. Comparison of chronic cough clusters demonstrated that certain combinations of medications and diagnoses characterize some chronic cough clusters.

Conclusion: To the best of our knowledge, this study is the first to test the potential of unsupervised deep learning methods for chronic cough investigation, which also shows a great advantage over existing algorithms for patient data clustering.

Identifying genetic markers enriched by brain imaging endophenotypes in Alzheimer's disease

Mansu Kim[1], Ruiming Wu[2], Xiaohui Yao[1], Andrew J. Saykin[3], Jason H. Moore1, Li Shen[1*] and for the Alzheimer's Disease Neuroimaging Initiative[4]

[1] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
[2] School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, USA.
[3] Indiana Alzheimer Disease Center and Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, USA.
[4] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[*] Corresponding author

Background: Alzheimer's Disease (AD) is a complex neurodegenerative disorder and the most common type of dementia. AD is characterized by a decline of cognitive function and brain atrophy, and is highly heritable with estimated heritability ranging from 60% to 80%. The most straightforward and widely used strategy to identify AD genetic basis is to perform genome-wide association study (GWAS) of the case-control diagnostic status. These GWAS studies have identified over 50 AD related susceptibility loci. Recently, imaging genetics has emerged as a new field where brain imaging measures are studied as quantitative traits (QTs) to detect genetic factors. Given that many imaging genetics studies did not involve the diagnostic outcome in the analysis, the identified imaging or genetic markers may not be related or specific to the disease outcome.

Results: We propose a novel method to identify disease-related genetic variants enriched by imaging endophenotypes, which are the imaging traits associated with both genetic factors and disease status.
Our analysis consists of three steps: 1) map the effects of a genetic variant (e.g., single nucleotide polymorphism or SNP) onto imaging traits across the brain using a linear regression model, 2) map the effects of a diagnosis phenotype onto imaging traits across the brain using a linear regression model, and 3) detect SNP-diagnosis association via

correlating the SNP effects with the diagnostic effects on the brain-wide imaging traits. We demonstrate the promise of our approach by applying it to the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Among 54 AD related susceptibility loci reported in prior large-scale AD GWAS, our approach identifies 41 of those from a much smaller study cohort while the standard association approaches identify only two of those. Clearly, the proposed imaging endophenotype enriched approach can reveal promising AD genetic variants undetectable using the traditional method.

Conclusion: We have proposed a novel method to identify AD genetic variants enriched by brain-wide imaging endophenotypes. This approach can not only boost detection power, but also reveal interesting biological pathways from genetic determinants to intermediate brain traits and to phenotypic AD outcomes.

AutoCoV: Tracking the Early Spread of COVID-19 in Terms of the Spatial and Temporal Dynamics from Embedding Space by K-mer Based Deep Learning

Inyoung Sung[1#,] Sangseon Lee[2#], Minwoo Pak[3], Yunyol Shin[3] and Sun Kim[1,3,4,5*]

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea.
[2] BK21 FOUR Intelligence Computing, Seoul National University, Seoul, Republic of Korea.
[3] Department of Computer Science and Engineering Seoul National University, Seoul, Republic of Korea.
[4] Institute of Engineering Research Seoul National University, Seoul, Republic of Korea.
[5] Bioinformatics Institute Seoul National University, Seoul, Republic of Korea.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: The widely spreading coronavirus disease (COVID-19) virus has three major properties: pathogenic mutations, spatial, and temporal propagation patterns. We know the spread of the virus geographically and temporally in terms of statistics, i.e., the number of patients. However, we are yet to understand the spread at the level of individual patients. As of March 2021, COVID-19 is wide-spread all over the world with new genetic variants. One important question is to track the early spreading patterns of COVID-19 until the virus has got spread all over the world.

Methods: In this work, we proposed a deep learning method, AutoCoV that can track the early spread of COVID-19 in terms of spatial and temporal dynamics of virus spreading patterns until the full spread over the world in July 2020. AutoCoV utilized information theoretic k-mer filtering to preprocess large genome sequences. Then, AutoCoV extended an auto-encoder network with a classifier module and a center loss objective function.

Results: Performances in learning spatial or temporal dynamics were measured with two clustering measures and one classification measure. For annotated SARS-CoV-2 sequences from the National Center for Biotechnology Information (NCBI), AutoCoV outperformed seven baseline methods in our experiments for learning either spatial or temporal dynamics. For spatial dynamics, AutoCoV had at least 1.7-fold higher clustering performances and an F1 score of 88.1%. For temporal dynamics, AutoCoV had at least 1.6-fold higher clustering performances and an F1 score of 76.1%. Furthermore, AutoCoV demonstrated the robustness of the embedding space with an independent dataset, Global Initiative for Sharing All Influenza Data (GISAID).

Conclusions: In summary, AutoCoV learns geographic and temporal spreading patterns successfully in experiments on NCBI and GISAID datasets and is the first of its kind that learns virus spreading patterns from the genome sequences, to the best of our knowledge. We expect that is type of embedding methods will be helpful in characterizing fast-evolving pandemics.

A framework to trace microbial engraftment at the strain level during FMT

Yiqi Jiang[1#], Shuai Wang[1#], Xianglilan Zhang[2*] and Shuaicheng Li[1*]

[1] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Ave, Kowloon Tong, Hong Kong, China.
[2] State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China.

[#] These authors contributed equally to this work
[*] Corresponding authors

Fecal microbiota transplantation (FMT) may treat microbiome-associated diseases effectively. However, the mechanism and pattern of the FMT process require expositions. Previous studies indicated the necessity to track the FMT process at the microbial strain level. At this moment, shotgun metagenomic sequencing enables us to study strain variations during the FMT.

We implemented a software package to study microbial strain variations during FMT from the shotgun metagenomic sequencing data. The package visualizes the strain alteration and traces the microbial engraftments during the FMT process. We applied the package to two typical FMT datasets, a ulcerative colitis (UC) dataset and a Clostridium difficile infection (CDI) dataset.

We observed that when the engrafted species has more than one strain in the source sample, 99.3% of the engrafted species will engraft only a subset of strains. We further confirmed that the all-or-nothing manner unsuited the engraftment of species with multiple strains by heterozygous SNPs count, revealing that strains prefer to engraft independently. Furthermore, we discovered a primary determinant of strain engrafted success is their proportion in species, as the donor engrafted strains and the pre-FMT engrafted strains with proportions 33.10% (p-value= 6e-06) and 37.0% (p-value = 9e-05) significantly higher than ungrafted strains on average, respectively. All the data sets indicated that the diversity of strains bursts after FMT and decreases to one after eight weeks for twelve species. Previous studies neglected strains with their corresponding species showing insignificant differences between different samples. With the package, from the UC data set, we successfully determined the strain variations of the species Roseburia intestinalis , a beneficial species reducing intestinal inflammation, colonized in the cured UC patient being engrafted from the donor, even if the patient hosted the same species yet before treatment; and from the CDI datasets, we found eight species that associated CDI FMT failure.

We demonstrated the necessity of analyzing whole-genome shotgun metagenomic data of FMT at the strain level. Also, we implemented a package to study FMT at the strain level and utilized it to uncover new knowledge about FMT. The package is available at https://github.com/deepomicslab/PStrain-tracer.

Exploration of Chemical Space with Partial Labeled Noisy Student Self-Training and Self-Supervised Graph Embedding: Application to Drug Metabolism

Yang Liu[1], Hansaim Lim[2], and Lei Xie[1*]

[1] Department of Computer Science, Hunter College, The City University of New York, 695 Park Ave, New York, NY 10065, United States.
[2] Ph.D. Program in Biochemistry, The Graduate Center, The City University of New York, 356
5th Ave, New York, NY 10016, United States.

[*] Corresponding author

Background: Drug discovery is time-consuming and costly. Machine learning, especially deep learning, shows great potential in accelerating the drug discovery process and reducing its cost. A big challenge in developing robust and generalizable deep learning models for drug design is the lack of a large amount of data with high-quality and balanced labels. To address this challenge, we developed a self-training method Partially LAbeled Noisy Student (PLANS) that uses a novel self-supervised graph embedding Graph-

Isomorphism-Network-based Fingerprint (GINFP) for chemical representations based on unlabeled data. PLANS-GINFP allows us to exploit millions of unlabeled chemical compounds as well as labeled and partially labeled pharmacological data to improve the generalizability of neural network models.

Results: We evaluated the performance of PLANS-GINFP for predicting Cytochrome P450 (CYP450) binding activity and chemical toxicity. The extensive benchmark studies proved that PLANS-GINFP could significantly improve the performance in both cases by a large margin. Both PLANS-based self-training and GINFP-based self-supervised learning contribute to performance improvement.

Conclusion: To better exploit chemical molecules as input for machine learning algorithms, we discovered a graph neural network-based embedding method that can convert chemical molecules to continuous-valued vectors. In addition to that, we developed a model agnostic method, PLANS, that can be applied to any deep learning architectures to improve prediction accuracies. PLANS also provided a way to better utilize partially labeled and unlabeled data. PLANS-GINFP could serve as a general solution to improve the predictive modeling for drug discovery.

Novel lincRNA Discovery and Tissue-Specific Gene Expression across 30 Normal Human Tissues

Xianfeng Chen[1] and Zhifu Sun[1*]

[1] Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic,
Rochester, MN 55905, USA.

[*] Corresponding author

Long non-coding RNAs (lncRNAs) are a large class of gene transcripts that do not code proteins; however, their functions are largely unknown and many new lncRNAs are yet to be discovered. Taking advantage of our previously developed, super-fast, novel lncRNA discovery pipeline, UClncR, and rich resources of GTEx RNA-seq data, we performed systematic novel lincRNA discovery for over 8000 samples across 30 tissue types. We conducted novel detection for each major tissue type first and then consolidated the novel discoveries from all tissue types. These novel lincRNs were profiled and analyzed along with known genes to identify tissue-specific genes in 30 major human tissue types. Thirteen sub-brain regions were also analyzed in a similar manner. Our analysis revealed thousands to tens of thousands of novel lincRNAs for each tissue type. These lincRNAs could define each tissue type's identity and demonstrated their reliability and tissue-specific expression. Tissue-specific genes were identified for each major tissue type and sub-brain region. The

tissue-specific genes clearly defined each respective tissue's unique function and could be used to expand the interpretation of non-coding SNPs from genome-wide association (GWAS) studies

---

Bioinformatics analysis revealed that functional important miRNA targeted genes are associated with Child Obesity trait in Genome-wide Association Studies

Melinda Song[1#], Jiaqi Yu[2#], Binze Li[3], Julian Dong[4], Jeslyn Gao[5], Lulu Shang[6], Xiang Zhou[6,7], Yongsheng Bai[8,9*]

[1] University of Michigan Medical School, Ann Arbor, MI 48109, USA.
[2] College Preparatory School, 6100 Broadway, Oakland, CA 94618, USA.
[3] Bellaire High School, 5100 Maple St, Bellaire, TX 77401, USA.
[4] Northville High School, 45700 Six Mile Road, Northville, MI 48168, USA.
[5] Simsbury High School, 34 Farms Village Rd, Simsbury, CT 06070, USA.
[6] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.
[7] Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA.
[8] Department of Biology, Eastern Michigan University, Ypsilanti, MI, 48197, USA.
[9] Next-Gen Intelligent Science Training, Ann Arbor, MI, 48105, USA.

[*] Corresponding author

Background: Genome Wide Association Studies (GWAS) have uncovered thousands of genetic variants that are associated with complex human traits and diseases. miRNAs are single-stranded non-coding RNAs. In particular, genetic variants located in the 3'UTR region of mRNAs may play an important role in gene regulation through their interaction with miRNAs. Existing studies have not been thoroughly conducted to elucidate 3'UTR variants discovered through GWAS. The goal of this study is to analyze patterns of GWAS functional variants located in 3'UTRs about their relevance in the network between hosting genes and targeting miRNAs, and elucidate the association between the genes harboring these variants and genetic traits.

Methods: We employed MIGWAS, ANNOVAR, MEME, and DAVID software packages to annotate the variants obtained from GWAS for 31 traits and elucidate the association between their harboring genes and their related traits. We identified variants that occurred in the motif regions that may be functionally important in affecting miRNA binding. We also conducted pathway analysis and functional annotation on miRNA targeted genes harboring 3'UTR variants for a trait with the highest percentage of 3'UTR variants occurring.

Results: The Child Obesity trait has the highest percentage of 3'UTR variants (75%). Of the 16 genes related to the Child Obesity trait, 5 genes (ETV7, GMEB1, NFIX, ZNF566,

ZBTB40) had a significant association with the term DNA-Binding ($p<0.05$). EQTL analysis revealed 2 relevant tissues and 10 targeted genes associated with the Child Obesity trait.
In addition, Red Blood Cells (RBC), Hemoglobin (HB), and Package Cell Volume (PCV) have overlapping variants. In particular, the PIM1 variant occurred inside the HB Motif region 37174641-37174660, and LUC7L3 variant occurred inside RBC Motif region 50753918-50753937.

Conclusion: Variants located in 3'UTR can alter the binding affinity of miRNA and impact gene regulation, thus warranting further annotation and analysis. We have developed a bioinformatics bash pipeline to automatically annotate variants, determine the number of variants in different categories for each given trait, and check common variants across different traits. This is a valuable tool to annotate a large number of GWAS result files.

MSPCD: Predicting circRNA-disease associations via integrating multi-source data and hierarchical neural network

Lei Deng[1], Dayun Liu[1], Yizhan Li[1], Runqi Wang[1], Junyi Liu[2], Jiaxuan Zhang[3] and Hui Liu[4*]

[1] School of Computer Science and Engineering, Central South University, 410083 Hunan, China.
[2] Viterbi School of Engineering, University of Southern California, 90089 Los Angeles, United States.
[3] Department of Cognitive Science, University of California San Diego, 92093 La Jolla, United States.
[4] School of Computer Science and Technology, Nanjing Tech University, 211816 Nanjing, China.

[*] Corresponding author

Background: Increasing evidence shows that circRNA plays an essential regulatory role in diseases through interactions with disease-related miRNAs. Identifying circRNA-disease associations is of great significance to precise diagnosis and treatment of diseases. However, the traditional biological experiment is usually time-consuming and expensive. Hence, it is necessary to develop a computational framework to infer unknown associations between circRNA and disease.

Results: In this work, we propose an efficient framework called MSPCD to infer unknown circRNA-disease associations. To obtain circRNA similarity and disease similarity accurately, MSPCD fifirst integrates more biological information such as circRNA-miRNA associations, circRNA-Gene Ontology (GO) associations, then extracts circRNA and disease high-order features by the neural network. Finally, MSPCD employs DNN to predict unknown circRNA-disease associations.

Conclusions: Experiment results show that MSPCD achieves a significantly more accurate performance compared with previous state-of-the-art methods on the circFunBase dataset. The case study also demonstrates that MSPCD is a promising tool that can effectively infer unknown circRNA-disease associations.

Gene co-expression changes underlying the functional connectomic alterations in Alzheimer's Disease

Bing He[1], Priyanka Gorijala[1], Linhui Xie[2], Sha Cao[3] and Jingwen Yan[1*]

[1] Department of BioHealth Informatics, Indiana University Purdue University Indianapolis, Indianapolis, Indiana, USA.
[2] Department of Electrical and Computer Engineering, Indiana University Purdue University Indianapolis, Indianapolis, Indiana, USA.
[3] Department of Biostatistics and Health Data Sciences, School of Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA.

[*] Corresponding author

There is growing evidence indicating that a number of functional connectivity networks are disrupted at each stage of the full clinical Alzheimer's disease spectrum. Such differences are also detectable in cognitive normal (CN) carrying mutations of AD risk genes, suggesting a substantial relationship between genetics and AD-altered functional brain networks. However, direct genetic effect on functional connectivity networks has not been measured. In this paper, we proposed a novel strategy to explore the genes related to altered functional brain connectivities in AD brains. Leveraging existing AD functional connectivity studies collected in NeuroSynth, we performed a meta-analysis to identify two sets of brain regions: ones with altered functional connectivity in resting state network and ones without. Then with the brain-wide gene expression data in the Allen Human Brain Atlas, we applied a new biclustering method to identify a set of genes with differential co-expression patterns between these two set of brain regions. We were able to identify 38 genes showing 4 modules of differential co-expression patterns.

Prioritization of risk genes in multiple sclerosis by a refined Bayesian framework followed by tissue-specificity and cell type feature assessment

Andi Liu[1#], Astrid M Manuel[2#], Yulin Dai[2], Zhongming Zhao[1,2,3*]

[1] Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA.
[2] Center for Precision Health, School of Biomedical Informatics, 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.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: Multiple sclerosis (MS) is a debilitating immune-mediated disease of the central nervous system that affects over 2 million people worldwide, resulting in a heavy burden to families and entire communities. Understanding the genetic basis underlying MS could help decipher the pathogenesis and shed light on MS treatment. We refined a recently developed Bayesian framework, Integrative Risk Gene Selector (iRIGS), to prioritize risk genes associated with MS by integrating the summary statistics from the largest GWAS to date (n = 115,803), various genomic features, and gene-gene closeness.

Results: We identified 163 MS-associated prioritized risk genes (MS-PRGenes) through the Bayesian framework. We replicated 35 MS-PRGenes through two-sample Mendelian randomization (2SMR) approach by integrating data from GWAS and Genotype-Tissue Expression (GTEx) expression quantitative trait loci (eQTL) of 19 tissues. We demonstrated that MS-PRGenes had more substantial deleterious effects and disease risk. Moreover, single-cell enrichment analysis indicated MS-PRGenes were more enriched in activated macrophages and microglia macrophages than non-activated ones in control samples. Biological and drug enrichment analyses highlighted inflammatory signaling pathways.

Conclusions: In summary, we predicted and validated a high-confidence MS risk gene set from diverse genomic, epigenomic, eQTL, single-cell, and drug data. The MS-PRGenes could further serve as a benchmark of MS GWAS risk genes for future validation or genetic studies.

SARS-COV-2 as potential microRNA sponge in COVID-19 patients

Chang Li[1*], Rebecca Wang[2], Aurora Wu[3], Tina Yuan[4], Kevin Song[5], Yongsheng Bai[6,7*] and Xiaoming Liu[1*]

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

[2] Pioneer High School, Ann Arbor, MI, USA.
[3] Emma Willard School, Troy, NY, USA.
[4] The Roeper School, Birmingham, MI, USA.
[5] Credit Suisse, New York, NY, USA.
[6] Next-Gen Intelligent Science Training, Ann Arbor, MI, USA.
[7] Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197, USA.

* Corresponding authors

Background: MicroRNAs (miRNAs) are a class of small non-coding RNA that can downregulate their targets by selectively binding to the 3' untranslated region (3'UTR) of most messenger RNAs (mRNAs) in the human genome. MiRNAs can interact with other molecules such as viruses and act as a mediator for viral infection. In this study, we examined whether, and to what extent, the SARS-CoV-2 virus can serve as a "sponge" for human miRNAs.

Results: We identified multiple potential miRNA/target pairs that may be disrupted during SARS-CoV-2 infection. Using miRNA expression profiles and RNA-Seq from published studies, we further identified a highly confident list of 5 miRNA/target pairs that could be disrupted by the virus's miRNA sponge effect, namely hsa-miR-374a-5p/APOL6, hsa-let-7f-1-3p/EIF4A2, hsa-miR-374a-3p/PARP11, hsa-miR-548d-3p/PSMA2 and hsa-miR-23b-3p/ZNFX1 pairs. Using single cell RNA sequencing based data, we identified hsa-miR-16-5p to be a potential virus targeting miRNA across multiple cell types from bronchoalveolar lavage fluid samples. We further validated some of our findings using miRNA and gene enrichment analyses and the results confirmed with findings from previous studies that some of these identified miRNA/target pairs are involved in pathways regulating pro-inflammatory cytokines and in immune cell maturation and differentiation.

Conclusion: Using publicly available databases and patient-related expression data, we found that acting as an "miRNA sponge" could be one explanation for SARS-CoV-2-mediated pathophysiological changes. This study provides a novel way of utilizing SARS-CoV-2 related data, with bioinformatics approaches, to help us better understand the etiology of the disease and its differential manifestation across individuals.

Bi-EB: An Empirical Bayesian Biclustering Algorithm for shared omics patterns between breast cancer tumor samples and breast cancer cell lines

Aida Yazdanparast[1,2,3], Lang Li[1,2,3,4*], Chi Zhang[1,2], and Lijun Cheng[4*]

[1] Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, IN, 46202.
[2] Department of Bio-Health Informatics, School of Informatics, Indiana University, Indianapolis, IN, 46202.
[3] Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis, IN, 46202.
[4] Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH, 43210.

[*] Corresponding authors

With the development of multi-omics data, it becomes urgent to monitor consensus module patterns of multi-omics data under various biological conditions, such as cancer cell lines and patients. A fast E mpirical B ayesian Bi clustering (Bi-EB) algorithm is developed to detect the local pattern of integrated multi-omics data across multiple conditions. Bi-EB adopts a data driven statistics strategy by using Expected-Maximum (EM) algorithm to extract the foreground bicluster pattern out of its background noise data in an iterative search. Bi-EB decomposed covariation from matrix row and column on a single gene level to seek common patterns by adjusting two parameters, bicluster membership probability threshold c and the bicluster average probability p. Our simulation experiments on three synthetic datasets display the Bi-EB model obtain high performance with 0.98 Recovery score and 0.99 Relevance scores comparing with seven popular biclustering methods, Cheng and Church (CC), xMOTIFs, BiMax, Plaid, Spectral, FABIA and QUBIC for constant, row and column shift-scaled biclustering searching. These methods obtain only 0.6 Recovery score and 0.5 Relevance score on average. The most importance is these methods cannot find a shared pattern across multiple conditions, nor seeking signaling transduction among multiple omics, such as from transcriptome to proteome. We utilized Bi-EB algorithm to determine shared pattern in mRNA and protein expression across subgroup of the Cancer Genomics Atlas (TCGA) and Cancer Cell Line Encyclopedia (CCLE) breast cancer samples. Transparent probabilistic interpretation and ratio strategy for omics data is first time proposed to detect the co-regulation patterns on protein and mRNA levels and identify their associated molecular functions. The largest bicluster has 12 gene/protein expression under 219 samples found, where include the clinically well-known over-expression genes estrogen receptor (ER) and ER (p118) and some novel genes

AR, BCL2, Cyclin E1 and IGFBP2 are recommended. Ten genes CCNB1, CDH1, KDR, RAB25, and PRKCA etc. are found which keep high accordance of mRNA/protein for both basal-like subtype in breast cancer patients and cell lines and become potential treatment targets for breast cancer.

Constrained tensor factorization for computational phenotyping and mortality prediction in patients with cancer

Francisco Y Cai[1], Chengsheng Mao[2], Yuan Luo[2*]

[1] Northwestern University Feinberg School of Medicine, Chicago 60611, IL, USA.
[2] Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago 60611, IL, USA.

[*] Corresponding author

Background: The increasing adoption of electronic health records (EHR) across the US has created troves of computable data, to which machine learning methods have been applied to extract useful insights. In particular, tensor factorization is one such method which has seen much development in recent years. EHR data, represented as a three-dimensional analogue of a matrix (tensor), is decomposed into two-dimensional factors that can be interpreted as computational phenotypes. Constraints imposed during the factorization can promote the discovery of phenotypes with certain desirable properties.

Methods: We apply constrained tensor factorization to derive computational phenotypes and predict mortality in cohorts of patients with breast, prostate, colorectal, or lung cancer in the Northwestern Medicine Enterprise Data Warehouse from 2000 to 2015. In our experiments, we examined using a supervised term in the factorization algorithm, filtering tensor co-occurrences by medical indication, and incorporating additional social determinants of health (SDOH) covariates in the factorization process. We evaluated the resulting computational phenotypes qualitatively and by assessing their ability to predict five-year mortality using the area under the curve (AUC) statistic.

Results: Filtering by medical indication led to more concise and interpretable phenotypes. Mortality prediction performance (AUC) varied under the different experimental conditions and by cancer type (breast: 0.623 – 0.694, prostate: 0.603 – 0.750, colorectal: 0.523 – 0.641, and lung: 0.517 – 0.623). Generally, prediction performance improved with the use of a supervised term and the incorporation of SDOH covariates.

Conclusion: Constrained tensor factorization, applied to sparse EHR data of patients with cancer, can discover computational phenotypes predictive of five-year mortality. The

incorporation of SDOH variables into the factorization algorithm is an easy-to-implement and effective way to improve prediction performance.

Small Molecule Modulation of Microbiota: A Systems Pharmacology Perspective

Qiao Liu[1], Bohyun Lee[2], Lei Xie[1,2,3,4*]

[1] Department of Computer Science, Hunter College, The City University of New York.
[2] Ph.D. Program in Computer Science, The City University of New York.
[3] Ph.D. Program in Biochemistry and Biology, The City University of New York.
[4] Helen and Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University.

[*] Corresponding author

Background: Microbes are associated with many human diseases and influence drug efficacy. Small-molecule drugs may revolutionize biomedicine by fine-tuning the microbiota on the basis of individual patient microbiome signatures. However, emerging endeavors in small-molecule microbiome drug discovery continue to follow a conventional "one-drug-one-target-one-disease" process. A systematic pharmacology approach that would suppress multiple interacting pathogenic species in the microbiome, could offer an attractive alternative solution.

Results: We construct a disease-centric signed microbe-microbe interaction network using curated microbe metabolite information and their effects on host. We develop a Signed Random Walk with Restart algorithm for the accurate prediction of effect of microbes on human health and diseases. With a survey on the druggable and evolutionary space of microbe proteins, we find that 8-10% of them can be targeted by existing drugs or drug-like chemicals and that 25% of them have homologs to human proteins. We demonstrate that drugs for diabetes can be the pioneer compounds for development of microbiota-targeted therapeutics. We further show that the potential drug targets that specifically exist in pathogenic microbes are periplasmic and cellular outer membrane proteins.

Conclusion: The systematic studies of polypharmacological landscape of the microbiome network may open a new avenue for the small-molecule drug discovery of microbiome. We believe that the application of systematic method on polypharmacological investigation could lead to the discovery of novel drug therapies.

Cost-effective Low-coverage Whole Genome Sequencing Assay for Glioma Risk Stratification

Jin Fu[1], Yingfeng Zhu[2], Xiaofeng Li[3], Jiajun Qin[1], Zhongrong Chen[1], Ziliang Qian[4,5], Jiping Sun[1*], Xianzhen Chen[1*]

[1] Department of Neurosurgery, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China.
[2] Department of Pathology, Huashan Hospital North, School of Medicine, Fudan University, Shanghai, China.
[3] Department of Emergency, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China.
[4] Suzhou Hongyuan Biotech Inc, Biobay, Suzhou, 215123.
[5] Prophet Genomics Inc, San Jose, CA, 95131.

[*] Corresponding authors

Current molecular pathological indicators of gliomas are cost intensive and complicated by multiple parameters. Here we investigated chromosomal instability (CIN), assayed with cost-effective low-coverage whole genome sequencing (WGS), as a biomarker for glioma risk stratifications. Thirty-five Formalin-Fixed Paraffin-Embedded gliomas samples were collected from Huashan Hospital. DNA was sent for WGS by Illumina X10 at low (median) genome coverage of 1.86x (range: 1.03-3.17x), followed by copy number analyses through a customized bioinformatics workflow Ultrasensitive Copy number Aberration Detector. As a result, we found chromosomal instability as a prognosis factor for gliomas independent of tumor grades. Patients with CIN+/7p11.2+ (12 grade IV and 3 grade III) had the worst survival (hazard ratio:16.2, 95% CI:6.3-41.6) with a median overall survival of 24 months. CIN+ patients without 7p11.2+ (6 grade III, 3 grade II) had a median survival of 65 months. Grade III patients with 7p11+ were distinguished at high risk of death as compared to the ones without 7p11 gain (Hazard ratio=16.2, P=0.0031). In summary, it is feasible to use cost-effective low-coverage WGS for risk stratification for glioma. Elevated chromosomal instability is associated with poor prognosis.

The profiling of gut microbiota during colorectal cancer development

Jingjing Liu[1, 2#], Wei Dong[1#], Jian Zhao[1], Jing Wu[3], Jinqiang Xia[2], Shaofei Xie[2], Xiaofeng Song[1*]

[1] Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China.
[2] The State Key Laboratory of Translational Medicine and Innovative Drug Development, Jiangsu Simcere pharmaceutical Co., Ltd., Nanjing 210016, China.
[3] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, Jiangsu, 211166, China.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: The imbalance of intestinal flora may promote the occurrence and development of colorectal cancer, changes of the intestinal flora during the development of colorectal cancer and the mechanism that promotes the colorectal cancer were discovered in this study. Deep sequencing of the microbial 16s ribosomal RNA gene was used to investigate alterations in feces samples of mice at the early inflammation stage and fully developed stage of colorectal cancer.

Results: According to PCoA analysis and ANOSIM test, we found the intestinal flora had significantly changed in mice with colorectal inflammation or colorectal cancer compared with healthy mice ($p<0.05$). Using correlation analysis, we found that S24-7 and Bacteroidaceae had strong excluding interactions. The functional changes of the gut microbiota include the up-regulation of the cancers pathway and the down-regulation of the replication and repair pathways.

Conclusion: Our study found the intestinal flora of mice suffering from colorectal inflammation and colorectal cancer has changed significantly, especially the decrease of S24-7 and the increase of Bacteroidaceae. We suppose that these two floras may play an important role in development of colorectal cancer.

Model performance and interpretability of semi-supervised generative adversarial networks to predict oncogenic variants with unlabeled data

Zilin Ren[1], Quan Li[1,2], Kajia Cao[3], Marilyn M Li[3,4], Yunyun Zhou[1*], Kai Wang[1, 4*]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.
[2] Princess Margaret Cancer Centre, University Health Network, University of Toronto, Toronto, Ontario, M5G2C1, Canada.
[3] Division of Genomic Diagnostics, Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA.
[4] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

[*] Corresponding authors

Background: It remains an important challenge to predict the functional consequences or clinical impacts of genetic variants in human diseases, such as cancer. An increasing number of genetic variants in cancer have been discovered and documented in public databases such as COSMIC, but the vast majority of them have no functional or clinical annotations. Some databases, such as CiVIC are available with manual annotation of functional mutations, but the size of the database is small due to the use of human annotation. Since the unlabeled data (millions of variants) typically outnumber labeled data (thousands of variants), computational tools that take advantage of unlabelled data may improve prediction accuracy.

Result: To leverage unlabeled data to predict functional importance of genetic variants, we introduced a method using semi-supervised generative adversarial networks (SGAN), incorporating features from both labeled and unlabeled data. Our SGAN model incorporated features from clinical guidelines and predictive scores from other computational tools. We also performed comparative analysis to study factors that influence prediction accuracy, such as using different algorithms, types of features, and training sample size, to provide more insights into variant prioritization. We found that SGAN can achieve competitive performances with small labeled training samples by incorporating unlabeled samples, which is a unique advantage compared to traditional machine learning methods. We also found that manually curated samples can achieve a more stable predictive performance than publicly available datasets.

Conclusions: By incorporating much larger samples of unlabeled data, the SGAN method can improve the ability to detect novel oncogenic variants, compared to other machine-learning algorithms that use only labeled datasets. SGAN can be potentially used to predict the pathogenicity of more complex variants such as structural variants or non- coding variants, with the availability of more training samples and informative features.

Copy Number Variation of Urine Exfoliated Cells s by Low-Coverage Whole Genome Sequencing for Diagnosis of Prostate Adenocarcinoma: A Prospective Cohort Study

Youyan Guan[1], Xiaobing Wang[2], Kaopeng Guan[1], Dong Wang[1], Xingang Bi[1], Zhendong Xiao[1], Zejun Xiao[1], Xingli Shan[3], Linjun Hu[3], Jianhui Ma[1], Changling Li[1], Yong Zhang[1], Jianzhong Shou[1], Baiyun Wang[4], Ziliang Qian[4*], Nianzeng Xing[1*]

[1] Department of Urology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, 100021, Beijing, China.
[2] State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, 100021 Beijing, China.
[3] Cancer Hospital of Huanxing, ChaoYang District, 100122, Beijing, China.
[4] Suzhou Hongyuan Biotech Inc, Biobay, Suzhou, 215123.

* Corresponding authors

Background: Non-invasive, urine-based diagnosis of prostate cancer (PCa) remains challenging. Although prostate cancer antigen (PSA) is widely used in prostate cancer screening, the false positives may result in unnecessary invasive procedures. PSA elevated patients are triaged to further evaluation of free/total PSA ratio (f/t PSA), to find out potential clinically significant PCa before undergoing invasive procedures. Genomic instability, especially chromosomal copy number variations (CNVs) were proved much more tumor specific. Here we performed a prospective study to evaluate the diagnostic value of CNV via urine-exfoliated cell DNA analysis in PCa.

Methods: We enrolled 28 PSA elevated patients (>=4 ng/ml), including 16 PCa, 9 Benign Prostate Hypertrophy (BPH) and 3 Prostatic Intraepithelial Neoplasia (PIN). Fresh initial portion urine was collected after hospital admission. Urine exfoliated cell DNA was analyzed by low coverage Whole Genome Sequencing, followed by CNV genotyping by the prostate cancer chromosomal aneuploidy detector (ProCAD). CNVs were quantified in absolute z-score (|Z|). Serum free/total PSA ratio (f/t PSA) was reported altogether.

Results: In patients with PCa, the most frequent CNV events were chr3q gain (n=2), chr8q gain (n=2), chr2q loss (n=4), and chr18q loss (n=3). CNVs were found in 81.2% (95% Confidence Interval (CI): 53.7-95.0%) PCa. No CNV was identified in BPH patients. A diagnosis model was established by incorporating all CNVs. At the optimal cutoff of |Z| ≥ 2.50, the model reached an AUC of 0.91 (95% CI: 0.83-0.99), a sensitivity of 81.2% and a specificity of 100%. The CNV approach significantly outperformed f/t PSA (AUC=0.62, P = 0.012). urther analyses showed that the CNV positive rate was significantly correlated with tumor grade. CNVs were found in 90.9% (95% CI: 57.1-99.5%) high grade tumors and 60.0% (95% CI: 17.0-92.7%) low grade tumors. No statistical significance was found for patient age, BMI, disease history and family history.

Conclusions: Urine exfoliated cells harbor enriched CNV features in PCa patients. Urine detection of CNV might be a biomarker for PCa diagnosis, especially in terms of the clinically significant high-grade tumors.

Association of the Tissue Infiltrated and Peripheral Blood Immune Cell Subsets with Response to Radiotherapy for Rectal Cancer

Min Zhu[1,3,4#], Xingjie Li[1,3#], Xu Cheng[3#], Xingxu Yi[3], Fang Ye[3], Xiaolai Li[4], Zongtao Hu[3], Liwei Zhang[3*], Jinfu Nie[1,3*], Xueling Li[1,3*]

[1] 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.
[2] Institute of Physical Science and Information Technology, Anhui University, 111 Jiulong Road, 230601, People's Republic of China.
[3] Hefei Cancer Hospital, Chinese Academy of Sciences.
[4] Hefei Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei 230031, People's Republic of China.

[#] These authors contributed equally to this work
[*] Corresponding authors

Background: Tumor microenvironment plays pivotal roles in carcinogenesis, cancer development and metastasis. Composition of cancer immune cell subsets can inferred by deconvolution of gene expression profile accurately. Compositions of the cell types in cancer microenvironment including cancer infiltrating immune and stromal cells have been reported to be associated with the cancer outcomes as markers for cancer prognosis. However, rare studies have been reported their association with the response to preoperative radiotherapy for rectal cancer.

Methods: In this paper, we deconvoluted the immune/stromal cell composition from the gene expression profiles. We compared the composition of immune/stromal cell types in the responsive vs. nonresponsive to RT for rectal cancer. We also analyzed the peripheral blood immune cell subset composition with fluorescence-activated cell sorting, and compared the immune cell subsets in the stable diseases vs. progressive diseases of rectal cancer patients from our institution.

Results: Compared with the non-responsive group, the responsive group showed higher proportions of CD4 + T cell (0.1378±0.0368 vs. 0.1071±0.0373, p = 0.0215), adipocytes, T cells CD4 memory resting, and lower proportions of CD8 + T cell (0.1798±0.0217 vs. 0.2104±0.0415, p = 0.0239), macrophages M2, and preadipocytes in their cancer tissue. The responsive patients showed a higher ratio of CD4+/CD8 + T cell proportions (mean 0.7869 vs. 0.5564, p = 0.0210). We also imported these eight cell features including eosinophils and macrophage M1 to Support Vector Machines and could predict the pre-radiotherapy responsive vs. non-responsive with an accuracy of 76%, AUC 0.77, 95% confidential interval of 0.632 to 0.857, better than the gene signatures. The results were further proved by a pooled dataset of GSE3493 and GSE35452. Consistently, the peripheral blood dataset also showed the higher proportion of CD4 + T cells and higher ratio of CD4 + /CD8 + T cells, lower proportion of CD8 + T cells for favorable prognosis.

Conclusions: Our results showed that both the proportions of tumor-infiltrating subsets and peripheral blood immune cell subsets can be important immune cell markers for outcomes of radiotherapy for rectal cancer treatment targets.

Comparison of four supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer

Tapas Bhadra[1#], Saurav Mallik[2#], Neaj Hasan[1] and Zhongming Zhao[2,3*]

[1] Department of Computer Science & Engineering, Aliah University, Kolkata, West Bengal 700160, India.
[2] Center for Precision Health, School of Biomedical Informatics, 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.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: As many complex omics data have been generated during the last two decades, dimensionality reduction problem has been a challenging issue in better mining such data. The omics data typically consists of many features. Accordingly, many feature selection algorithms have been developed. The performance of those feature selection methods often varies by specific data, making the discovery and interpretation of results challenging.

Methods and results: In this study, we performed a comprehensive comparative study of five widely used supervised feature selection methods (mRMR, INMIFS, DFS, SVM-RFE-CBR and VWMRmR) for multi-omics datasets. Specifically, we used five representative datasets: gene expression (Exp), exon expression (ExpExon), DNA methylation (hMethyl27), copy number variation (Gistic2), and pathway activity dataset (Paradigm IPLs) from a multi-omics study of acute myeloid leukemia (LAML) from The Cancer Genome Atlas (TCGA). The different feature subsets selected by the aforesaid five different feature selection algorithms are assessed using three evaluation criteria: (i) classification accuracy (Acc), (ii) representation entropy (RE) and (iii) redundancy rate (RR). Four different classifiers, viz., C4.5, NaiveBayes, KNN, and AdaBoost, were used to measure the classification accuary (Acc) for each selected feature subset. The VWMRmR algorithm obtains the best Acc for four datasets (Exp, ExpExon, hMethyl27 and Paradigm IPLs). The VWMRmR algorithm offers the best RR (obtained using normalized mutual information) for three datasets (Exp, Gistic2 and Paradigm IPLs), while it gives the best RR (obtained using Pearson correlation coefficient) for two datasets (Gistic2 and Paradigm IPLs). It also obtains the best RE for three datasets (Exp, Gistic2 and Paradigm IPLs). Overall, the VWMRmR algorithm yields best performance for all three evaluation criteria for majority of the datasets. In addition, we identified signature genes using supervised learning collected from the overlapped top feature set among five

feature selection methods. We obtained a 7-gene signature for EXP, a 9-gene signature for ExpExon, a 7-gene signature for hMethyl27, one single-gene signature (P IK3CG) for Gistic2 and a 3-gene signature for Paradigm IPLs.

Conclusion: We performed a comparison of performance evaluation of five feature selection methods for mining features from various high-dimensional datasets and then identified signature genes.

A Landscape of Immune Cell Types in Tumor Microenvironment Associated with Prognosis and Sensitivity of Radiotherapy

Xingjie Li[1,2], Min Zhu[2,3*], Xueling Li[2,3*]

[1] Institute of Physical Science and Information Technology, Anhui University, 111
Jiulong Road, 230601.
[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.
[3] Hefei Cancer Hospital, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei 230031, People’s Republic of China, 350 Shushanhu Road, Hefei 230031.

[*] Corresponding authors

Background: Radiotherapy is a regular treatment for tumors, but only part of patients have desired response. Recent studies show that the effectiveness of treatment can be influenced by the tumor immune microenvironment. Therefore, it is very important to predict the outcome of the treatment accurately for patients to get the maximum benefit from radiotherapy.

Results: By using Wilcoxon test and univariate Cox proportional hazards regression model, respectively, we analyzed the data of the 11 cancers with radiotherapy from The Cancer Genome Atlas, and investigated the relationship between the compositions of infiltrating immune cell subsets and radiotherapy prognosis in terms of complete response or not, overall survival and recurrence free survival. We found that some statistically significant cell types are consistent in two or more cancers, including CD4 naïve T cells, Pericytesas favorable factors and, Plasma cells as unfavorable. In overall survival and recurrence free survival univariate Cox proportional hazards regression analysis, however, Brain Lower Grade Glioma show opposite trends in terms of cell type composition with Head and Neck squamous cell carcinoma for radiotherapy. Higher compositions of B cell, dendritic, neutrophil, mast cells resting, and T cells regulatory are unfavorable for lower grade glioma, while they are all favorable prognosis factors for head and neck squamous cell carcinoma. Mast cell activated shows opposite.. We found that iDC in Cervical

squamous cell carcinoma and endocervical adenocarcinoma, B cells and fibroblasts in Head and Neck squamous cell carcinoma, T cells CD4 naïve in Brain Lower Grade Glioma are consistently favorable in OS, RFS and CR vs. ICR. In addition, we established a survival prognosis model based on multivariate Cox proportional hazards regression for brain lower grade glioma based on the cell type compositions in the tumor microenvironment.

Conclusions: We provide a landscape of the immune cell types associated with the RT responses and prognosis in eleven cancers. We confirmed that the role of immune cells depends on cancers and treatment endpoints.

Revealing the novel complexity of plant long non-coding RNA by strand-specific and whole transcriptome sequencing for evolutionarily representative plant species

Yan Zhu[1#], Longxian Chen[1,2#], Xiangna Hong[1,3], Han Shi[1,2], Xuan Li[1,2*]

[1] Key Laboratory of Synthetic Biology, Center for Excellence in Molecular Plant Sciences/Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China.
[2] University of Chinese Academy of Sciences, Beijing, China.
[3] Henan University, Kaifeng, China.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: Previous studies on plant long noncoding RNAs (lncRNAs) lacked consistency and suffered from many factors like heterogeneous data sources and experimental protocols, different plant tissues, inconsistent bioinformatics pipelines, etc. For example, the sequencing of RNAs with poly(A) tails excluded a large portion of lncRNAs without poly(A), and use of regular RNA-sequencing technique did not distinguish transcripts' direction for lncRNAs. The current study was designed to systematically discover and analyze lncRNAs across eight evolutionarily representative plant species, using strand-specific (directional) and whole transcriptome sequencing (RiboMinus) technique.

Results: A total of 39,926 lncRNAs (25,331 lincRNAs and 14,595 lncNATs) were identified, which showed molecular features of lncRNAs that are consistent across divergent plant species but different from those of mRNA. Further, transposable elements (TEs) were found to play key roles in the origination of lncRNA, as significantly large number of lncRNAs were found to contain TEs in gene body and promoter region, and transcription of many lncRNAs was driven by TE promoters. The lncRNA sequences were divergent even in closely related species, and most plant lncRNAs were genus/species-specific, amid rapid turnover in evolution. Evaluated with PhastCons scores, plant lncRNAs showed similar conservation level to that of intergenic sequences, suggesting that most lincRNAs were young and with short evolutionary age. INDUCED BY PHOSPHATE STARVATION (IPS) was found so far to be the only plant lncRNA group with conserved motifs, which may play important roles in the adaptation of terrestrial life during migration from aquatic to terrestrial. Most highly and specially expressed lncRNAs formed co-expression network with coding genes, and their functions were believed to be closely related to their co-expression genes.

Conclusion: The study revealed novel features and complexity of lncRNAs in plants through systematic analysis, providing important insights into the origination and evolution of plant lncRNAs.

Identifying Alzheimer’s Genes via Brain Transcriptome Mapping

Jae Young Baik[1#], Mansu Kim[2#], Jingxuan Bao[1], Qi Long[2], Li Shen[2*] and for the Alzheimer's Disease Neuroimaging Initiative[3]

[1] School of Arts and Sciences, University of Pennsylvania, Philadelphia, USA.
[2] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
[3] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[#] These authors contributed equally to this work
[*] Corresponding author

Background: Alzheimer’s disease (AD) is one of the most common neurodegenerative disorders characterized by progressive decline in cognitive function. Targeted genetic analyses, genome-wide association studies, and imaging genetic analyses have been performed to detect AD risk and protective genes and have successfully identified dozens of AD susceptibility loci. Recently, brain imaging transcriptomics analyses have also been conducted to investigate the relationship between neuroimaging traits and gene expression measures to identify interesting gene-traits associations. These imaging transcriptomic studies typically do not involve the disease outcome in the analysis, and thus the identified brain or transcriptomic markers may not be related or specific to the disease outcome.

Results: We propose an innovative two-stage approach to identify genes whose expression profiles are related to diagnosis phenotype via brain transcriptome mapping. Specifically, we first map the effects of a diagnosis phenotype onto imaging traits across the brain using a linear regression model. Then, the gene-diagnosis association is assessed by spatially correlating the brain transcriptome map with the diagnostic effect map on the brain-wide imaging traits. To demonstrate the promise of our approach, we apply it to the integrative analysis of the brain transcriptome data from the Allen Human Brain Atlas (AHBA) and the amyloid imaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Our method identifies 12 genes whose brain-wide transcriptome patterns are highly correlated with six different diagnostic effect maps on the amyloid imaging traits. These 12 genes include four confirmatory findings (i.e., AD genes reported in DisGeNET) and eight novel genes that have not be associated with AD in DisGeNET.

Conclusion: We have proposed a novel disease-related brain transcriptomic mapping method to identify genes whose expression profiles spatially correlated with regional diagnostic effects on a studied brain trait. Our empirical study on the AHBA and ADNI data shows the promise of the approach, and the resulting AD gene discoveries provide valuable information for better understanding biological pathways from transcriptomic signatures to intermediate brain traits and to phenotypic disease outcomes.

On triangular inequalities of correlation-based distances for gene expression profiles

Jiaxing Chen[1], Yen Kaow Ng[2], Lu Lin[1], Xianglilan Zhang[3*] and Shuaicheng Li[1*]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong,.
[2] Department of Computer Science, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia.
[3] State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, 100071, Beijing, People’s Republic of China.

[*] Corresponding authors

Distance functions are fundamental for evaluating the differences between gene expression profiles. Such a function would output a low value if the profiles are strongly correlated—either negatively or positively—and vice versa. One popular distance function is the absolute correlation distance, da = 1 − |ρ|, where ρ is similarity measure, such as Pearson or Spearman correlation. However, the absolute correlation distance fails to fulfill the triangle inequality, which would have guaranteed better performance at vector quantization, allowed fast data localization, as well as accelerated data clustering. In this work, we propose dr = □1 − |ρ| as an alternative. We prove that dr satisfies the triangular equality when ρ represents. Pearson correlation, Spearman correlation, or Cosine □ similarity. We show dr to be better than ds = that satisfies the triangle inequality, both analytically as well as experimentally. We empirically compared dr with da in gene clustering and sample clustering experiment by real-world biological data. The two distances performed similarly in both gene clusters and sample clusters in hierarchical clustering and PAM (partitioning around medoids) clustering. However, dr demonstrated more robust clustering. According to the bootstrap experiment, dr generated more robust sample pair partition more frequently (p-value < 0.05). The class “dissolved” event also support the advantage in robustness.

Mouse Blood Cells Types and Aging Prediction using Penalized Latent Dirichlet Allocation

Xiaotian Wu[1], Yee Voan Teo[2], Nicola Neretti[2] and Zhijin Wu[1*]

[1] Department of Biostatistics, Brown University, Providence, RI, US.
[2] Department of Molecular Biology, Cell Biolgy, and Biochemistry, Brown University, Providence, RI, US.

[*] Corresponding author

Aging is a complex, heterogeneous process that has multiple causes. Knowledge on genomic, epigenomic and transcriptomic changes during the aging process shed light on understanding the aging mechanism. A recent breakthrough in biotechnology, single cell RNAseq, is revolutionizing aging study by providing gene expression profile of the entire transcriptome of individual cells. Many interesting information could be inferred from this new type of data with the help of novel computational methods. In this manuscript a novel statistical method, penalized Latent Dirichlet Allocation (pLDA), is applied to an aging mouse blood scRNA-seq data set. A pipeline is built for cell type and aging prediction. The sequence of models in the pipeline take scRNAseq expression counts as input, preprocess the data using pLDA and predict the cell type and aging status.

Rewired Pathways and Disrupted Pathway Crosstalk in Schizophrenia Transcriptomes by Multiple Differential Coexpression Methods

Hui Yu[1], Yan Guo[1], Jingchun Chen[2], Xiangning Chen[2], Peilin Jia[3] and Zhongming Zhao [3,4,5*]

[1] Department of Internal Medicine, University of New Mexico, Albuquerque, NM 87131, USA.
[2] Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, Las Vegas, NV 89154, USA.
[3] Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA.
[4] Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA.
[5] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA.

[*] Corresponding author

Transcriptomic studies of mental disorders using the human brain tissues have been limited, and gene expression signatures in schizophrenia (SCZ) remain elusive. In this study, we applied three differential co-expression methods to analyze five transcriptomic datasets (three RNA-Seq and two microarray datasets) derived from SCZ and matched normal postmortem brain samples. We aimed to uncover biological pathways where internal correlation structure was rewired or inter-coordination was disrupted in SCZ. In total, we identified 60 rewired pathways, many of which were related to neurotransmitter, synapse, immune, and cell adhesion. We found the hub genes, which were on the center of rewired pathways, were highly mutually consistent among the five datasets. The combinatory list of 92 hub genes was generally multi-functional, suggesting their complex and dynamic roles in SCZ pathophysiology. In our constructed pathway crosstalk network, we found "Clostridium neurotoxicity" and "signaling events mediated by focal adhesion kinase" had the highest interactions. We further identified disconnected gene links underlying the disrupted pathway crosstalk. Among them, four gene pairs (PAK1:SYT1, PAK1:RFC5, DCTN1:STX1A, and GRIA1:MAP2K4) were normally correlated in universal contexts. In summary, we systematically identified rewired pathways, disrupted pathway crosstalk circuits, and critical genes and gene links in schizophrenia transcriptomes.

kESVR: An ensemble model for drug response prediction in precision medicine using cancer cell lines gene expression

Abhishek Majumdar[1], Yueze Liu[2], Yaoqin Lu[3], Shaofeng Wu[1] and Lijun Cheng[1*]

[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA.
[2] The Grainger College of Engineering, The University of Illinois Urbana-Champaign, Urbana and Champaign, Champaign, IL 61801, USA.
[3] Department of Occupational and Environmental Health, School of Public Health, Xinjiang Medical University, Urumqi 830011, China.

[*] Corresponding author

Background: Cancer cell lines are frequently used in research as in-vitro tumor models. Genomic data and large-scale drug screening have accelerated the right drug selection for cancer patients. Accuracy in drug response prediction is crucial for success. Due to data-type diversity and big data volume, few methods can integrative and efficiently find the principal low-dimensional manifold of the high-dimensional cancer multi-omics data to predict drug response in precision medicine.

Method: A novelty k-means Ensemble Support Vector Regression (kESVR) is developed to predict each drug response values for single patient based on cell-line gene expression data. The kESVR is a blend of supervised and unsupervised learning methods and is entirely data driven. It utilizes embedded clustering (Principal Component Analysis and k-means clustering) and local regression (Support Vector Regression) to predict drug response and obtain the global pattern while overcoming missing data and outliers' noise.

Results: We compared the efficiency and accuracy of kESVR to 4 standard machine learning regression models: (1) simple linear regression, (2) support vector regression (3) random forest (quantile regression forest) and (4) back propagation neural network. Our results, which based on drug response across 610 cancer cells from Cancer Cell Line Encyclopedia (CCLE) and Cancer Therapeutics Response Portal (CTRP v2), proved to have the highest accuracy (smallest mean squared error (MSE) measure). We next compared kESVR with existing 17 drug response prediction models based a varied range of methods such as regression, Bayesian inference, matrix factorization and deep learning. After ranking the 18 models based on their accuracy of prediction, kESVR ranks first (best performing) in majority (74%) of the time. As for the remaining (26%) cases, kESVR still ranked in the top five performing models.

Conclusion: In this paper we introduce a novel model (kESVR) for drug response prediction using high dimensional cell-line gene expression data. This model outperforms current existing prediction models in terms of prediction accuracy and speed and overcomes overfitting. This can be used in future to develop a robust drug response prediction system for cancer patients using the cancer cell-lines guidance and multi-omics data.

Revealing the hadal viral community in the sediment of New Britain Trench

Hui Zhou[1,2#], Ping Chen[1#], Mengjie Zhang[1,2], Jiawang Chen[3*], Jiasong Fang[4*] and Xuan Li[1,2*]

[1] Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences,
Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai 200032, China.
[2] University of Chinese Academy of Sciences, Beijing 100049, China.
[3] Ocean College, Zhejiang University, Zhoushan 316021, China.
[4] Shanghai Engineering Research Center of Hadal Science and Technology, College of Marine Sciences, Shanghai Ocean University, Shanghai 201306, China.

[#] These authors contributed equally to this work
[*] Corresponding authors

Marine viruses are widely distributed and influence matter and energy transformation in ecosystems by modulating hosts' metabolism. The hadal trenches represent the deepest marine habitat on Earth, for which the viral communities and related biogeochemical functions are least explored and poorly understood. Here, using the sediment samples (8720 m below sea level) collected from the New Britain Trench (NBT), we investigated the viral community, diversity, and genetic potentials in the hadal sediment habitat for the first time by deep shotgun metagenomic sequencing. We found the NBT sediment viral community was dominated by Siphoviridae, Myoviridae, Podoviridae, Mimiviridae, and Phycodnaviridae, which belong to the dsDNA viruses. However, the large majority of them remained uncharacterized. We found the hadal sediment virome had some common components by comparing the hadal sediment viruses with those of hadal aquatic habitats and those of bathypelagic and terrestrial habitats. It was also distinctive in community structure and had many novel viral clusters not associated with the other habitual virome included in our analyses. Further phylogenetic analysis on its Caudovirales showed novel diversities, including new clades specially evolved in the hadal sediment habitat. Annotation of the NBT sediment viruses indicated the viruses might influence microbial hydrocarbon biodegradation and carbon and sulfur cycling via metabolic augmentation through auxiliary metabolic genes (AMGs). Our study filled in the knowledge gaps on the virome of the hadal sediment habitats and provided insight into the evolution and the potential metabolic functions of the hadal sediment virome.

Alternative splicing induces sample-level variation in gene-gene correlations

Yihao Lu[1], Brandon L. Pierce[1,2], Pei Wang[3], Fan Yang[4] and Lin S. Chen[1*]

[1] Department of Public Health Sciences, University of Chicago, 5841 South Maryland Ave MC2000, Chicago, IL 60637.
[2] Department of Human Genetics, University of Chicago, 920 E 58th St, Chicago, IL 60637.
[3] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 770 Lexington Ave, New York, NY 10065.
[4] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, 13001 E. 17th Place, Aurora, Colorado 80045.

[*] Corresponding author

The vast majority of genes in the genome are multi-exonic, and are alternatively spliced during transcription, resulting in multiple isoforms for each gene. For some genes, different mRNA isoforms may have differential expression levels or be involved in different pathways. Bulk tissue RNA-seq, as a widely used technology for transcriptome quantification, measures the total expression (TE) levels of each gene across multiple

isoforms in multiple cell types for each tissue sample. With recent developments in precise quantification of alternative splicing
events for each gene, we propose to study the effects of alternative splicing variation on gene-gene correlation effects. We adopted a variance-component model for testing the TE-TE correlations of one gene with a co-expressed gene, accounting for the effects of splicing variation and splicing-by-TE interaction of one gene on the other. By analyzing data from the Genotype-Tissue Expression (GTEx) project (V8), we showed that splicing variation of a gene may interact
with TE of the gene and affect the TE of co-expressed genes, resulting in substantial inter-sample variability in gene-gene correlation effects. At the 5% FDR level, 38,146 pairs of genes out of ~10M examined pairs from GTEx lung tissue showed significant TE-splicing interaction effects, implying isoform-specific and/or sample-specific TE-TE correlations. Additional analysis across 13 GTEx brain tissues revealed strong tissue-specificity of TE-splicing interaction effects.
Moreover, we showed that accounting for splicing variation across samples could improve the reproducibility of results and could reduce potential confounding effects in studying co-expressed gene pairs with bulk tissue data. Many of those gene pairs had correlation effects specific to only certain isoforms and would otherwise be undetected. By analyzing gene-gene co-expression variation within functional pathways accounting for splicing, we characterized the patterns of the "hub" genes with isoform-specific regulatory effects on multiple other genes.

APA-Scan: Detection and Visualization of 3'-UTR APA with RNA-seq and 3'-end-seq Data

Naima Ahmed Fahmi[1], Khandakar Tanvir Ahmed[1], Jae-Woong Chang[3], Heba Nassereddeen[2],
DeliangFan[4], Jeongsik Yong[3†] and Wei Zhang[1*†]

[1] Department of Computer Science, University of Central Florida, 4000 Central Florida Blvd, Orlando, Florida, 32816, United States.
[2] Department of Computer Engineering, University of Central Florida, 4000 Central Florida Blvd, Orlando, Florida, 32816, United States.
[3] Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota Twin Cities, 420 Washington Ave. S.E., Minneapolis, MN 55455, United States.
[4] School of Electrical, Computer and Energy Engineering, Arizona State University, 650 E Tyler Mall, Tempe, AZ 85287, United States.

[*] Corresponding author

Background: The eukaryotic genome is capable of producing multiple isoforms from a gene by alternative polyadenylation (APA) during pre-mRNA processing. APA in the 3'-untranslated region (3'-UTR) of mRNA produces transcripts with shorter or longer 3'-UTR. Often, 3'-UTR serves as a binding platform for microRNAs and RNA-binding proteins, which affect the fate of the mRNA transcript. Thus, 3'-UTR APA is known to modulate translation and provides a mean to regulate gene expression at the post-transcriptional level. Current bioinformatics pipelines have limited capability in profiling 3'-UTR APA events due to incomplete annotations and a low-resolution analyzing power: widely available bioinformatics pipelines do not reference actionable polyadenylation (cleavage) sites but simulate 3'-UTR APA only using RNA-seq read coverage, causing false positive identifications.

Results: To overcome these limitations, we developed APA-Scan, a robust program that identifies 3'-UTR APA events and visualizes the RNA-seq short-read coverage with gene annotations. APA-Scan utilizes either predicted or experimentally validated actionable polyadenylation signals as a reference for polyadenylation sites and calculates the quantity of long and short 3'-UTR transcripts in the RNA-seq data. APA-Scan works in three major steps: (A) calculate the read coverage of the 3'-UTR region of genes; (B) identify the potential APA sites and evaluate the significance of the events among two biological conditions; (C) graphical representation of user specific event with 3'-UTR annotation and read coverage on the 3'-UTR region. APA-Scan is implemented in Python3. Source code and a comprehensive user's manual are freely available at https://github.com/compbiolabucf/APA-Scan.

Conclusion: APA-Scan was applied to both simulated and real RNA-seq datasets and compared with two widely used baselines DaPars and APAtrap. In simulation APA-Scan significantly improved the accuracy of 3'-UTR APA identification compared to the other baselines. The performance of APA-Scan was also validated by 3'-end-seq data and qPCR on mouse embryonic fibroblast cells. The experiments confirm that APA-Scan can detect unannotated 3'-UTR APA events and improve genome annotation. APA-Scan is a comprehensive computational pipeline to detect transcriptome-wide 3'-UTR APA events. The pipeline integrates both RNA-seq and 3'-end-seq data information and can efficiently identify the significant events with a high-resolution read-coverage plots.

Poster - Abstracts

Abstract 1

Computational Saturation Mutagenesis of Coronavirus Proteins

Shaolei Teng[1], Adebiyi Sobitan[1], Vidhyanand Mahase[1], Raina Rhoades[1], Dongxiao Liu[2], Qiyi Tang*[2]

[1]Department of Biology, Howard University, Washington DC, 20059 USA
[2]Howard University College of Medicine, Washington DC, 20059 USA

*Corresponding Author

The human pathogenic coronaviruses, including severe acute respiratory syndrome coronavirus (SARS-CoV-1), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and Middle East respiratory syndrome coronavirus (MERS-CoV), are serious threats to global health. The current pandemic of the COVID19 caused by SARS-CoV-2 has devastatingly affected almost all countries with more than 179 million infected cases and 3.8 million deaths to date. No proven effective therapeutic is available, and it will likely come back next season. Coronaviruses are positive-sense RNA viruses that can easily generate mutations as viruses spread, which is the major challenge for the ongoing development of broad neutralizing antibodies. The RNA genomics of coronaviruses generate various non-structural proteins and structural proteins including spike (S) proteins. SARS-CoV-1 and SARS-CoV-2 S proteins bind to the host receptors including angiotensin converting enzyme 2 (ACE2). The bioinformatics methods can readily quantify the effects of mutations on protein functions and structures. We have used computational saturation mutagenesis approaches to quantify the systemic effects of missense mutations on S proteins of SARS-CoV-2, SARS-CoV-1 and MERS-CoV and their human receptors on protein stability and binding affinity. The structure-based free energy calculation methods and sequence-based machine learning tools are used to quantify the effects of missense mutations on protein stability, virus-receptor interaction and post translational modifications. The findings provide useful information for characterizing the functional effects of key mutations on protein structure and function. The knowledge gained from this research can be used to investigate other virus proteins and help scientists to design drugs for the next pandemic.

Abstract 2

Converting Tabular Data into Images for Anti-Cancer Drug Response Prediction Using Convolutional Neural Networks

Yitan Zhu[1], Thomas Brettin[1], Fangfang Xia[1], Alexander Partin[1], Maulik Shukla[1], Hyunseung Yoo[1], Yvonne A. Evrard[2], James H. Doroshow[3], Rick L. Stevens*[1,4]

[1]Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL 60439, USA
[2]Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc. Frederick, MD 21702, USA
[3]Developmental Therapeutics Branch, National Cancer Institute, Bethesda, MD 20892, USA;[4]Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA

*Corresponding Author

Convolutional neural networks (CNNs) have been successfully used in many applications where important information about data is embedded in the order of features, such as speech and imaging. However, most tabular data do not assume a spatial relationship between features, and thus are unsuitable for modeling using CNNs. To meet this challenge, we develop the image generator for tabular data (IGTD), a novel algorithm that transforms tabular data into images by assigning features to pixel positions so that similar features are close to each other in the image. The algorithm searches for an optimized assignment by minimizing the difference between the ranking of distances between features and the ranking of distances between their assigned pixels in the image. We apply IGTD to transform gene expression profiles of cancer cell lines (CCLs) and molecular descriptors of drugs into their respective image representations for the application of anti-cancer drug response prediction. Compared with existing methods for converting tabular data into images, IGTD possesses multiple advantages. First, it does not require priori knowledge about the relationship between features and thus can be used in the absence of domain knowledge. Second, it generates compact image representations, in which each pixel represents a unique feature. Compact image representations can help reduce memory consumption and model training time in the subsequent prediction modeling using CNNs. Third, the IGTD image representations better preserve the feature neighborhood structure. Features that are closely located in the IGTD images are indeed more similar. Fourth, evaluated on benchmark drug screening datasets, CNNs trained on IGTD image representations of CCLs and drugs exhibit a better performance of predicting anti-cancer drug response than both CNNs trained on alternative image representations and prediction models trained on the original tabular data. Fifth, IGTD provides a flexible framework to accommodate diversified data and requirements. Various distance measures can be used to calculate the feature and pixel distances. The number of dimensions, size, and shape of the images can be flexibly chosen. The IGTD framework can be extended in a straightforward manner to transform data vectors into not only 2-D matrices, but also 1-D or multi-dimensional arrays or even images of irregular shapes, such as a concave polygon. A full paper about this work has been recently published. The reference information is Y. Zhu et

al., Converting tabular data into images for deep learning with convolutional neural networks, *Scientific Reports*, vol. 11, article number: 11325 (2021)

Abstract 3

Bioinformatics analysis of gene expression profiles to identify key proteins associated with Parkinson's disease

Baharak Akbaria[1] Kamran Rafiei[1], Najaf Allahyari Fard*[1]

[1] Departement of systems biotechnology, National institute of Genetic engineering and biotechnology (NIGEB), Tehran, Tehran, Iran

*Corresponding Author

Parkinson's disease is the second most common disease of the nervous system after Alzheimer's disease and is characterized by the gradual loss of dopaminergic neurons in the substantia nigra. The annual incidence of this disease is estimated at 18-18 per 100,000 people. Studies have shown that the incidence and prevalence of this disease increases with age, so that the prevalence of this disease in the total population is about 0.3% and in the population over 60 years is about 1%. Due to the increase in the average age of the population in the world, it is thought that the prevalence of this disease is progressing rapidly. The aim of this study was to identify the key genes, pathways involved and compare the expression pattern of the identified genes in the incidence of Parkinson's disease. Raw microarray data with access code GSE7621 was obtained from Geo database. Geo 2R online software was used to identify genes with different expression (Geo). Gene cluster enrichment analysis was performed using the Metascape website. STRING database and network analysis software including Cytoscape 3.8.2 and CentiScaPe 2.2 were used to reach the gene pathways involved in Parkinson's disease, plotting the protein protein interaction (PPI) network and reaching key genes. The results of data analysis indicate that several genes increased and several genes decreased and also several genes changed between the healthy and diseased groups. After drawing the protein interaction network, GRIN1, DRD2, CXCR4 ESR1, KNG1 and PTPRC genes were identified as key genes for Parkinson's disease. The present study suggests that some of the major genes and pathways may be associated with the development of Parkinson's disease and may also be considered as important biomarkers for new therapeutic propos in the future.

Abstract 4

Machine Learning-based Identification of Molecular Signatures of Sex-biased Genes for Breast Cancer Survival

Eric W. Li[1], Yongsheng Bai*[2]

[1]Lakeside School, 14050 1st Ave NE, Seattle, WA 98125, USA
[2]Department of Biology, Eastern Michigan University, 441 Mark Jefferson Hall,
Ypsilanti, MI 48197, USA

*Corresponding Author

The role of microRNAs (miRNAs), which perform their functions through targeting messenger RNAs (mRNAs), are often altered when aberrant expression is present. In addition, X chromosome-located (X-linked) miRNAs have a broad role in cell lineage determination, immune regulation, and oncogenesis. Sex-biased genes could contribute to sex bias in cancer due to expression change by targeting miRNAs. How biological roles and associations with immune cell abundance levels for sex-biased gene-miRNA pairs in gender-related cancer (e.g., breast cancer) change due to the alteration of their expression pattern to identify candidate therapeutic markers has not been investigated thoroughly. Upon analyzing anti-correlated genes and miRNAs within significant clusters of 12 The Cancer Genome Atlas (TCGA) cancer types and the list of sex-biased genes and miRNAs reported from previous studies, 125 sex-biased genes (11 male-biased and 114 female-biased) were identified in breast cancer (BC). Seventy-three sex-biased miRNAs (40 male-biased and 33 female-biased) were identified across 5 out of 12 cancers (head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), and lung adenocarcinoma (LUAD)). Correlation between the BC sex-biased genes and tumor infiltrating immune cell types was further evaluated. Machine Learning algorithms were also used to elucidate factors contributing to overall patient survival. Specifically, the multivariate Cox regression analysis was performed to identify a gene expression molecular signature significantly correlated to patient overall survival time. We found eight genes having high correlation with immune infiltration. Fifteen candidate female-biased BC genes targeted by 3 X-linked miRNAs (*hsa-mir-18, hsa-mir-221*, and *hsa-mir-224*) were pinpointed in this study. The molecular signature CD2-low, SCML4-low, CXCR6-high, and CD8B-low was identified to have significantly worse patient survival time (p=0.02). Our computational result indicates that many identified female-biased genes which have positive associations with immune cell abundance levels could serve as alternative therapeutic markers. Our analysis suggests that female-biased expression of BC candidate genes is likely influenced by their targeting miRNA(s). We plan to investigate immune response over post-transcriptional regulatory network in the context of breast cancer etiology for different racial groups as well as for different subtypes, such as luminal A, luminal B, and triple negative BC.

Abstract 5

Functional Screening of 3’-UTR Variants Combined with Genome-wide Association Identifies Causal Regulatory Mechanisms Impacting Alcohol Consumption

Andy B Chen[1,2], Kriti S. Thapa[3], Hongyu Gao[1,2,4], Jill L Reiter[1,3], Hongmei Gu[3], Junjie Zhang[1], Xiaoling Xuei[1,4], Dongbing Lai[1], Yue Wang[1], Howard J. Edenberg[1,3], Yunlong Liu*[1,2,4]

[1]Department of Medical & Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA
[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA
[3]Department of Biochemistry & Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA
[4]Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN, USA

*Corresponding Author

Genome-wide association studies (GWAS) have identified loci associated with alcohol consumption. Many are in non-coding regions, and additional information is needed to identify the functional elements within the loci. Using a massively parallel reporter assay (MPRA) in neuroblastoma and microglia cell lines, we evaluated the functional activity of variants in the 3’ untranslated regions (3’-UTR) of genes associated with neurological disorders. Of the 13,515 SNPs tested, 400 and 657 significantly impacted gene expression in neuroblastoma and microglia, respectively. We then analyzed the heritability enrichment of the functionally relevant variants identified by MPRA in two GWAS of alcohol use: drinks per week from GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN); and alcohol use disorder (AUD) from the Million Veteran Program (MVP). We found that functionally active variants explained a higher proportion of heritability than other variants tested in both cell lines. We integrated the estimated effects from MPRA results with GWAS results from GSCAN drinks per week and MVP alcohol use disorders identification test-consumption (AUDIT-C) to identify genes in which the activity of SNPs in the 3’-UTR were associated with alcohol consumption. We found 7 genes in neuroblastoma and 6 genes in microglia that were replicated in both GWAS. For each gene, we calculated an MPRA-based 3’-UTR activity level using the genotypes of brain tissue samples from the CommonMind Consortium. The top and bottom third of samples based on the activity level were separated and the genes differentially expressed between them were identified, yielding 1,160 genes from neuroblastoma and 2,159 genes from microglia. A pathway analysis of these differentially expressed genes identified several inflammation response pathways, including IL-17, NF-kappa B, and TNF signaling. Because these genes were identified by their genetic components, these results suggest that variation in response to inflammation may be a causal factor in differences in alcohol consumption. In this study, integrating MPRA and GWAS results allowed us to gain insight into genetic mechanisms

affecting alcohol consumption, and using this framework can help elucidate these relationships for genes within GWAS loci in other traits.

Abstract 6

Contribution of transposable elements to tissue-specific gene regulation in human

Arsala Ali, Ping Liang*[1]

[1]Department of Biological Sciences, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, L2S 3A1, Ontario, Canada

*Corresponding Author

Transposable elements (TEs) have been known to play a regulatory role in the host genome. With varying epigenomic profile of TEs across different tissue types, TE-derived regulatory sites tend to be tissue-specific via TE-derived tissue-specific promoters and enhancers. We aim to determine and compare genes (and their associated biological processes) enriched for TE-derived regulatory sites in the cell lines of different tissues, based on regulatory regions demarcated by active and repressed chromatin states. The cell lines analyzed were DND-41 (Blood), GM12878 (Blood), GM23248 (skin/limb), HCT116 (Colon), HeLa-S3 (Cervix), HepG2 (Liver), IMR-90 (Lung), Karpas-422 (Blood), MCF-7 (Breast), MM.1S (Blood), NCI-H929 (Bone marrow), PC-3 (Prostate), PC-9 (Lung), SK-N-SH (Brain), chosen based on data availability for all DNase-seq and histone ChIP-seq experiments. DNase-seq data was recruited from ENCODE data portal. Comparing pattern of TEs in shared DNase hypersensitive sites (DHS) and cell line specific DHS (appeared in one or more but not all cell lines) showed that LTRs make 44% of total TEs in shared DHS vs. 29% in cell line specific DHS, and it is 26% vs. 36%, 13% vs. 23% and 17% vs. 11% for LINEs, SINEs and DNAs, respectively. We thus found over-representation of LTRs and DNAs while under-representation of LINEs and SINEs in shared DHS. We further collected genes with ≥10% of DHS in gene neighboring region (5Kb upstream and 5Kb downstream of transcription start site (TSS)) being TE-derived and performed GO enrichment analysis. For all cell lines except for MCF-7, GM12878, GM23248 and SK-N-SH, enrichment for different biological processes were found, with some showing cell-specific patterns. For example, biological processes including nucleosome organization, chromatin assembly/disassembly, defense response to virus and DNA conformation change were found only for DND41 (blood). Further, comparing hierarchical clustering of cell lines based on presence/absence of DHS and TE-derived DHS (TE-DHS) in genes showed very similar clustering patterns indicative of their tissue origin, confirming the important role of TEs in tissue-specific gene regulation. Overall, our results suggest that all major types of TEs in the human genome contribute to gene regulation in a tissue- or

cell type-specific pattern. We further aim to extend the analysis to other gene regulatory datasets (histone active sites and histone repressive sites) to get a broad picture regarding tissue-specificity of TE-contributed gene regulation.

Abstract 7

Analysis of factors impacting the quality of *de novo* genome assemblies

Haimeng (Jerry) Tang[1], Ping Liang[1], Adonis Skandalis[1], Miriam Richards*[1]

[1]Department of Biological Sciences, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, L2S 3A1, Ontario, Canada

*Corresponding Author

Next generation sequencing (NGS) technology has revolutionized genomic and genetic research, and as a result, *de novo* sequencing and genome assembly for non-model organisms has now become a common pursuit in genomic research. However, the individual properties of genomes, such as ploidy, sequence variation, and repeats, impose challenges for genome assemblers. In this study, we use *Xylocopa virginica* (Eastern carpenter bees) as a unique model organism for examining the effects of haplodiploidy (males are haploid, females are diploid), sequence variability, repeat content, coverage, and sequencing platform on genome assembly quality. We assembled four bee genomes for comparisons by age and sex (unworn male, worn male, unworn female, and worn female (worn bees being older)) using SoapDenovo2 with standard Illumina sequencing (coverage from 134X to 163X). We also assembled another unworn female genome using Supernova with 10X linked-reads sequencing (coverage at 250X) for comparison by platform. We discovered noticeable differences in assembly quality metrics among these genome assemblies. Specifically, we found that the haploid, unworn male genome had the highest overall quality and the diploid, worn female genome had the lowest quality, with the N50 value of the former being more than 100 times higher than that of the latter. Furthermore, the density of variants was moderately correlated to the density of contig/scaffold breakpoints. The pattern of genome quality supports the hypothesis that genetic heterozygosity resulting both from ploidy and somatic variants can affect the quality of an assembly, with ploidy playing a much larger role. With Illumina short reads from 5X to 40X sequence coverage, the unworn male showed the biggest genome quality increase followed by the worn male, unworn and worn female assemblies, while for 10X linked-reads for the diploid genome, sequencing at 63X coverage seems to offer the best genome quality. Altogether, our results showed that sequencing depth has variable effects on genome assembly quality depending on the diploidy and age of the animals, as well as the sequence platform. In conclusion, our results indicate that for a *de novo* assembly project for a non-model organism, the use of haploid samples at the youngest possible age and

sequencing at high coverage, preferably with a long-read platform, can achieve the best genome assembly. However, at least for the 10X linked-reads, having more sequencing coverage beyond a certain threshold does not necessarily lead to a better genome assembly.

Abstract 8

A machine-learning classifier for predicting aneuploidy risk in female IVF patients

Siqi Sun,[1] Maximilian Miller,[2] Yanran Wang,[2] Katarzyna M. Tyc,[1,6] Richard T. Scott, Jr.,[3] Xin Tao,[4] Yana Bromberg,[1,2,5] Karen Schindler,[1,5] Jinchuan Xing*[1,5]

[1] Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA.
[2] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ USA
[3] Reproductive Medicine Associates of New Jersey, Basking Ridge, NJ, USA.
[4] Foundation for Embryonic Competence, Basking Ridge, NJ, USA.
[5] Human Genetic Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ, USA.
[6] VCU Massey Cancer Center, Richmond, VA, USA.

*Corresponding Author

Infertility is a major reproductive health issue that affects about 8.8% of women in the United States. Aneuploidy is the leading genetic cause of infertility in women of reproductive age, in which the embryo has abnormal numbers of chromosomes. Recent studies show that genetic variants in several genes decrease the fidelity of chromosome segregation and predispose women to a higher incidence of egg aneuploidy. However, the exact genetic cause of aneuploid egg production remains unknown, making it difficult to diagnose infertile patients and suggest reproductive choices based on the presence of certain germline variants. In this study, we developed a machine learning (ML)-based classifier to predict the risk of embryonic aneuploidy in female patients (categorized as either high or low risk) and identify candidate genes that contribute to the aneuploidy phenotype. Using whole-exome sequencing data of *in vitro* fertilization (IVF) patient samples, we constructed a Random Forest ML model to predict the embryonic aneuploidy risk status of patients, whose data was not used in model development. In two independent datasets, we achieved the highest area under receiver operating curve (ROC-AUC) of 0.77 and 0.68, respectively. High precision or high specificity can be achieved to classify patients by selecting different prediction score cutoffs. For example, using a strict prediction score cutoff of 0.7, we can identify 29% of high-risk patients with 94% precision. Among the genes that contribute the most to the predictive power of the model, we found potential aneuploidy causing genes (e.g., *MCM5*). In summary, using two exome-

sequencing datasets, we demonstrated the feasibility of developing a classifier for screening aneuploidy risk in female IVF patients. The classification results can potentially provide guidance on individualized treatment options, including more accurate prognosis of IVF treatment success, number of required IVF cycles, or recommendations for alternative family planning options. Patients with lower prediction scores might have higher chances of successful IVF treatment, whereas patients with higher prediction scores would be suggested to have multiple cycles of IVF treatments at a younger age or adopt alternative family plans.

Abstract 9

DRAGOM: Classification and Quantification of Noncoding RNA in Metagenomic Data

Ben Liu[1], Sirisha Thippabhotla[1], Jun Zhang[2,3], Cuncong Zhong*[1,4,5,*]

[1]Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS, United States
[2]Division of Medical Oncology, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, United States
[3]Department of Cancer Biology, University of Kansas Medical Center, Kansas City, KS, United States
[4]Bioengineering Program, The University of Kansas, Lawrence, KS, United States
[5]Center for Computational Biology, The University of Kansas, Lawrence, KS, United States.

[*]Corresponding Author

Noncoding RNAs (ncRNAs) play important regulatory and functional roles in microorganisms, such as regulation of gene expression, signaling, protein synthesis, and RNA processing. Hence, their classification and quantification are central tasks toward the understanding of the function of the microbial community. However, the majority of the current metagenomic sequencing technologies generate short reads, which may contain only a partial secondary structure that complicates ncRNA homology detection. Meanwhile, de novo assembly of the metagenomic sequencing data remains challenging for complex communities. To tackle these challenges, we developed a novel algorithm called DRAGoM (Detection of RNA using Assembly Graph from Metagenomic data). DRAGoM first constructs a hybrid graph by merging an assembly string graph and an assembly de Bruijn graph. Then, it classifies paths in the hybrid graph and their constituent reads into different ncRNA families based on both sequence and structural homology. We benchmarked DRAGoM with the read-based strategy (homology search against unassembled reads) and the assembly-based strategy (homology search against assembled contigs) on subsampled CAMI (the Critical Assessment of Metagenome Interpretation) challenge dataset and subsampled real human gut microbiome dataset. On the CAMI dataset, DRAGoM achieved the best F-score (79.3% to 92.8%) when searching different non-16S rRNA families. It also achieved the best F-score (96.4%) when searching 16S rRNA. On the real dataset (SRR341583), DRAGoM achieved the best F-score (74.4%) for 60 non-16S rRNA families, and comparable 16S rRNA search performance. Our benchmark experiments show that DRAGoM can improve the performance and robustness over traditional approaches on the classification and quantification of a wide class of ncRNA families.

Abstract 10

iMPP: Integrated De Novo Gene Prediction for Metagenomics Data

Sirisha Thippabhotla[1], Ben Liu[1], Cuncong Zhong*[1,2,3]

[1]Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS, United States
[2]Bioengineering Program, The University of Kansas, Lawrence, KS, United States
[3]Center for Computational Biology, The University of Kansas, Lawrence, KS, United States.

*Corresponding Author

Abstract

Metagenomics analysis allows the exploration of the genomic content of microbial communities residing in different environmental niches. De novo prediction of protein-coding genes from the resulted genomic sequence is a key step towards novel gene family discovery and understanding of the function of microbial communities. De novo gene prediction on metagenomic data is challenging because of the difficulty of reconstructing complete microbial genomes, due to the high taxonomic diversity of the microbial communities and limitations of current sequencing technologies. We have developed a pipeline called iMPP (the integrated Metagenomic Protein Predictor), which leverages sequence overlap information, to improve the stability and prediction accuracy of the current open reading frame (ORF) based prediction models. Specifically, iMPP has three key components: a core de novo gene prediction engine, a nucleotide assembly engine for constructing and traversing a hybrid nucleotide assembly graph (constructed by merging string graph and *de Bruijn* graph) and a peptide assembly engine. The peptide assembly stage helps to recruit any protein-coding reads that may have been missed in the former prediction stages. We benchmarked the performance of iMPP against traditional read-based approaches and assembly-based approaches. We tested our pipeline on both simulated and subsampled real datasets. Overall, iMPP was able to push the F-scores for de novo gene prediction to 92-93%, from an already-high ground of 84-87% on the subsampled real datasets. iMPP's improved performance further benefited downstream de novo peptide assembly, by generating more assembled protein sequences (3.46-6.75%), along with improved N50 (0-1.35%) and total contig lengths (1.02-4.10%). The protein contigs were also able to cover more reference protein sequences at varying sequencing length cutoffs (0.98-3.39%). Our benchmarking results suggest that iMPP will have important applications in identifying and reconstructing novel protein sequences from metagenomics data, a critical step towards unbiased and comprehensive understanding of the functional aspects of microbial community.

Abstract 11

Bayesian Mixed-Effect Higher-Order Hidden Markov Models with Applications to Predictive Healthcare Using Electronic Health Record

Ying Liao[1], Yisha Xiang[1], Zhigen Zhao[2], Di Ai*[3]

[1]Department of Industrial, Manufacturing & Systems Engineering, Texas Tech University, Lubbock, TX, USA
[2]Department of Statistical Science, Fox School of Business Temple University, Philadelphia, PA, USA
[3]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

*Corresponding Author

Predictive modeling using electronic health records (EHRs) offers great promise for making informed clinical decisions and improving healthcare quality. The EHR data generally exist significant variability due to the mixed effects of some critical exogenous factors (for example, gender and age) and patient-level heterogeneity. Ignoring these effects may lead to inaccurate prediction results and further cause inferior treatments. We develop a novel and flexible Bayesian mixed-effect higher-order hidden Markov model (MHOHMM) to incorporate fixed and random effects, and propose an MHOHMM-based framework for clinical prediction. We construct the MHOHMM in a Bayesian hierarchical formulation and design an effective Markov chain Monte Carlo (MCMC) sampling algorithm for statistical inference. A simulation study is conducted to evaluate the performance of the proposed sampling algorithm and the MHOHMM-based prediction method. Several model structures are designed to investigate the impacts of mixed effects on parameter estimation and sequence prediction. Our simulation results show that the proposed sampling algorithm is effective for MHOHMM inference and incorporation of mixed effects significantly improves prediction performance when these effects present in the data. The practical utility of the proposed predictive framework is demonstrated by a case study on the acute hypotensive episode (AHE) prediction for intensive care unit patients using the MIMIC-III (Medical Information Mart for Intensive Care) database. The

results show that the proposed method provides good prediction performance in clinical practices.

Abstract 12

Gene regulatory networks and biomarkers jointly inferred by genome-wide genetic, epigenetic, and regulatory factors in multiple sclerosis

Astrid M Manuel[1], Yulin Dai[1], Hyun-Hwan Jeong[1], Zhongming Zhao*[1,2,3]

[1]Center for Precision Health, The University of Texas Health Science Center at Houston, Houston, TA, USA
[2]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

*Corresponding Author

Growing biological big data provides unprecedented opportunities for systematically modeling the disease-context gene regulatory networks (GRNs), reflecting underlying ‘causal’ inferences. Multiple sclerosis (MS) is a heritable autoimmune disease featured with chronic neuronal injury and disability. Previous studies have demonstrated both genetic and epigenetic factors contribute to MS pathogenesis and relapse. In this study, we utilized the multi-omics data from peripheral blood of MS patients to explore the genetic and epigenetic impacts on their GRNs. The epigenomic dataset included DNA methylation profiles of peripheral blood from 279 MS cases and controls. We also integrated association signals of 8,868,766 single-nucleotide polymorphisms (SNPs) investigated in the largest genome-wide association study (GWAS) of MS, which assayed the genotypes of 14,802 MS cases and 26,703 controls. Furthermore, we used expression profiles from peripheral blood of 337 cases and controls. Here, MS multi-omics data were integrated by our in-house Edge-Weighted Dense Module Search (EW_dmGWAS) tool to acquire the prioritize gene set. This genetic- and epigenetic-enriched GRN was then reconstructed by applying a tree-based Random Forest method to predict the potential transcription factors. The Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining (TRRUST) database was used to validate the transcription factors and their targets relationships. The resultant GRN included 27 genes and 26 directed edges, depicting regulatory functions of genes in peripheral blood of MS patients. Importantly, two drug targets of MS FDA-approved medications were present in the GRN: *RELA* and *MS4A1*, which warrants further investigation.

Abstract 13

Functional annotation of mutations in the alpha, beta, gamma and delta variants of SARS-CoV-2

Theodore Jiang[1, 2, 3], Andy Wang[2], Qian Liu[3], Kai Wang* [3]

[1]Palisades Charter High School, Pacific Palisades, CA 90272, USA
[2]Princeton Middle School, Princeton, NJ 08540, USA
[3]Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

*Corresponding Author

Throughout the COVID-19 pandemic, new variants of the SARS-CoV-2 virus emerge, but it is unknown the specific mechanisms that allow them to take over regions of the world so quickly. We aim to analyze the nucleotide mutations and their corresponding protein sequence and structure changes to see what potential mutations may enable certain variants of concern (VOC) to dominate. We downloaded SARS-CoV-2 sequences from the GISAID database, identified mutations, and annotated their functional effects on proteins using ANNOVAR. We paid special attention to mutations on the spike protein of the Delta variant, and highlighted the locations of the mutations in 3D structure on the folded spike protein. We generated graphs of the frequency of each variant over time to see patterns of competition between variants over time. We are particularly interested in the T478K mutation of Delta variant's spike protein, so we utilized the mutagenesis function in PyMOL to visualize the specific molecular structural changes to the ACE2 binding domain. The Alpha, Beta, and Gamma variants have all decreased in cases while the Delta variant has increased during the month of May 2021 worldwide. Some mutations in Delta variant's spike protein may be the cause for this dominance: (1) T478K, part of the human ACE2 binding receptor (437-508aa), could be responsible for Delta variant's increased transmissibility. Upon 3D visualization, the T478K mutation reaches closer to position 23 (Glutamic acid) and 24 (Glutamine) on human ACE2, potentially forming a salt bridge with position 23. (2) P681R, part of the spike protein cleavage site (680-685aa), which affects furin-mediated spike cleavage, may explain the Delta variant's increased pathogenicity. (3) T19R, located in the N1 loop (14-20aa) of "supersite", where all known antibodies bind, and it may affect vaccine efficacy. (4) p.E156_R158del, a nonframeshift mutation located on the N3 loop of supersite (140-158aa), may also affect vaccine efficacy. Additionally, we found that D614G and L452R are both on important sites in the spike protein, but are not unique to Delta variant, so they are unlikely to be major causes for Delta variant's recent dominance over other VOC. The Delta variant (B1.617.2) significantly increased in prevalence recently compared to other variants of concern (Alpha, Beta, Gamma). The Delta variant has several unique mutations in its spike protein, including T478K, which could play an important role for Delta variant's recent dominance in parts of the world.

Charting the Proteome Landscape in Major Psychiatric Disorders: From Biomarkers to Biological Pathways

Brisa S. Fernandes [1], Yulin Dai [1], Peilin Jia [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 UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA

*Corresponding Author

Background: Schizophrenia (SZ), Bipolar Disorder (BD), and Major Depressive Disorder (MDD) share many of their underlying biologies and there is a lack of biological validity in their diagnosis, which is a hindrance to developing diagnostic tests. Proteomic studies have evolved the field by identifying several proteins that could be biomarkers in those disorders, however, their results have varied widely, and individual biomarkers have failed to advance diagnostics. Our goal is to leverage the broad variability among different studies and strengthen the knowledge provided by individual proteins by systematically interrogating the literature to uncover biological pathways with stronger biological meaning.Methods: This study is a systematic review of all proteomics studies in BD, MDD, and SZ in blood. We extracted all differentially expressed proteins in BD, MDD, or SZ. We then conducted ORA and GSEA to unveil which biological pathways were shared or unique to each disorder.

Results: We included 51 studies with 9423 participants. 486 proteins were found altered in cases compared to controls (192 in SZ; 190 in BD; and 365 in MDD). The majority of the enriched pathways were shared among SZ, BD, and MDD. The top pathways in all three disorders were associated with the immune system and complement cascade. Other pathways shared among SZ, BD, and MDD were interleukin-12 signaling, MAPK1/MAPK3 signaling, PI3K-Akt Signaling, Toll-like Receptor Signaling, Activation of Matrix Metalloproteinases, Class A/1 (Rhodopsin-like receptors), GPCR downstream signaling, Advanced glycosylation end-product receptor signaling, JAK-STAT signaling, and Regulation of Insulin-like Growth Factor (IGF) transport. Pathways shared between SZ and BD were integrin cell-surface interactions and syndecan interactions. Shared between BD and MDD were NRF2 pathway, signaling by EGFR, and Ras signaling pathway. Unique to SZ were interleukin receptor SHC signaling, and TFAP2 (AP-2) family

regulation of growth factors; unique to MDD were oncostatin M signaling, ECM-receptor interaction, plasminogen activating cascade, and PPAR signaling pathway.

Discussion: Alterations in pathways related to immune-inflammation were pervasive and transdiagnostic. That the immunoinflammatory response, as assessed in peripheral blood, is a shared construct across SZ, BD, and MDD might imply that the periphery is an unspecific representation of a mechanism placed in the brain and probably a secondary phenomenon to a primarily central origin and that the biological boundaries among SZ, BD, and MDD mostly do not resemble current nosological categories and need to be completely redefined by the identification of new sub-types that may or may not overlap with brain-derived sub-types.

Abstract 15

Single cell-based deconvolution of liver diseases reveals γδ2 T cells as a marker in hepatocellular carcinoma development

Rama Shankar[1], Mingdian Tan[3], Joseph W. Zagorski[4], Austin J. Goodyke[4], Jeremy Haskins[1], Shreya Paithankar[1], Dave Chesla[4], Samuel So[3], Mei-Sze Chua[3], Bin Chen*[1,2]

[1]Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI 49503, USA
[2]Department of Pharmacology and Toxicology, College of Human Medicine, Michigan State University, Grand Rapids, MI 49503, USA
[3]Asian Liver Center, Department of Surgery, School of Medicine, Stanford University, Stanford, California, 94305, USA
[4]Spectrum Health, Grand Rapids, MI 49503, USA

*Corresponding Author

Hepatocellular Carcinoma (HCC) is the third leading cause of cancer-related deaths worldwide. HCC morbidity is the highest in individuals with chronic liver diseases (CLDs); however, the effects of cell composition on the progression of CLDs to HCC remains unknown. Using gene biomarkers of 20 distinct cell types derived from healthy liver single cell RNA-seq data, we estimated the cell compositions of a total of six CLDs and HCC based on their RNA-seq profiles. We observed distinct changes of normal cell type enrichment landscape among liver fibrosis, non-alcoholic fatty liver disease, and HCC. Compared to the healthy state, liver fibrosis and HCC present higher enrichment of γδ2 T cells and lower enrichment of central venous liver sinusoidal endothelial cells in RNA-seq and scRNA-seq data. High enrichment of γδ2 T cells was predominant in HCC with underlying chronic hepatitis B or C virus infections, as well as in advanced HCC, and was associated with poor prognosis in patients with HCC, particularly those with underlying virus infection. The enrichment scores of γδ2 T cells remarkably classified dysplastic nodules, early-stage HCC, and advanced stage HCC from the healthy state (AUC: 0.87, 0.89, 0.98, respectively). Immunofluorescence staining of liver tissues confirmed a monotonic elevation of γδ2 T cells from healthy, inflammation, cirrhosis to HCC. Furthermore, fluorescence-activated cell sorting of PBMC samples revealed that more than 45% of HCC patients presented abnormally elevated level of γδ2 T cells circulating in the blood. Together, we identified γδ2 T cells as a novel and translational marker in HCC development.

Abstract 16

Simulating Double Minute Evolution using Java

Derrick Mullins[1], Matthew Hayes*[1]

[1]Department of Physics and Computer Science, Xavier University of Louisiana, New Orleans, LA, USA

*Corresponding Author

Double minutes are small fragments of circular DNA. Unlike typical chromosomes, they are composed of circular fragments of DNA, up to only a few million base pairs in size and contain no centromere or telomere. Double minutes highly amplified and formed as a byproduct of chromothripsis, or excision and circulation of genomic segments. They are known to harbor oncogenes (genes that are overexpressed) and cause cancer onset when overexpressed. This Java program simulates the evolution of double minutes using recursion, which is the repeated application calling itself. The number of replications are dependent upon the number of generations starting from the parent double minute chromosome. The double minutes are presented in BED format, which shows only the chromosome the double minute is located, along with the start and end chromosomes of the double minute.

Abstract 17

Brain age acceleration estimated from functional network connectivity as biomarker of Alzheimer’s disease progression

Mohammad. S. E. Sendi[1,2,3], David H Salat[4,5], Vince D Calhoun*[1,2,3,7]

[1]Wallace H. Coulter Department of Biomedical Engineering at Emory University and Georgia Tech, Atlanta, GA, USA
[2]School of Electrical and Computer Engineering (ECE) at the Georgia Institute of Technology, Atlanta, GA, USA

[3]Tri-Institutional Georgia State University/Georgia Institute of Technology/Emory University Center for Translational Research in Neuroimaging and Data Science, Atlanta, GA, USA [4]Neuroimaging Research for Veterans Center, VA Boston Healthcare System, Boston, MA, USA
[5]Massachusetts General Hospital, Charlestown, MA, USA
[6]Harvard Medical School, Boston, MA, USA, [7]Georgia State University, Atlanta, GA, USA

*Corresponding Author

Introduction: The brain age gap, the difference between an individual's brain predicted age and chronological age, is used as a brain disease biomarker and aging. To date, although previous studies utilized mostly structural magnetic resonance imaging (MRI) data to predict brain age, less work has used functional network connectivity (FNC) from resting-state functional MRI (rs-fMRI) to indicate brain age and its link with Alzheimer's disease (AD) progression. This study aims to estimate brain age from FNC and introduce it as a biomarker of AD progression.

Methods: This study used 1091 rs-fMRI data and the chronological age at the time of scanning, ranging from 42-95, from the Open Access Series of Imaging Studies (OASIS)-3 cohort. This dataset includes 1021 healthy subjects (HC) and 70 AD patients. We used group independent component analysis to estimate 53 components for the whole-brain. Then, FNC feature, in total 1378, of each subject was calculated by applying the Pearson correlation. A support vector regression (SVR) was trained on 951 HC subjects in which FNC features and the chronological age was served as predictor and prediction output, respectively. We then tested the trained model on 70 HC (age mean± sd: 73.66± 7.47) and 70 AD (age mean± sd: 73.34± 7.51) subjects to predict their brain age.

Results: The correlation between predicted brain age and the chronological age of HC test data was R=0.73, ($p=5.3e^{-13}$, N=70), while the correlation between predicted brain age and AD patients' chronological age was R=0.33 (p=0.004, N=70). This result shows that the model based on HC could better predict the HC age than the AD patients. The brain age gap mean and standard deviation for the test HC and AD group was -2.2581± 5.0879 and 2.0814± 7.4520. This result shows that the brain age gap was increased in AD (Cohen's d=0.68, p<0.001).

Conclusions: We proved that FNC, estimated from rs-fMRI, could predict the brain age in HC. Also, we found an acceleration in the brain predicted age of AD patients. We also found that the AD subjects' brain age gap was significantly higher than the brain age gap of the HC group

Abstract 18

Brain dynamic functional connectivity predicts treatment response to electroconvulsive therapy in major depressive disorder

Mohammad. S. E. Sendi[1,2,3], Hossein Dini[4], Christopher C. Abbott[5], Vince D Calhoun*[1,2,3,6]

[1]Wallace H. Coulter Department of Biomedical Engineering at Emory University and Georgia Tech, Atlanta, GA, USA
[2]School of Electrical and Computer Engineering (ECE) at the Georgia Institute of Technology, Atlanta, GA, USA
[3]Tri-Institutional Georgia State University/Georgia Institute of Technology/Emory University Center for Translational Research in Neuroimaging and Data Science, Atlanta, GA, USA, [4]Department of Architecture, Design and Media Technology, Aalborg University, Copenhagen, Denmark
[5]Department of Psychiatry, University of New Mexico, Albuquerque, NM, United States, [6]Georgia State University, Atlanta, GA, USA

*Corresponding Author

Background: Electroconvulsive Therapy (ECT) is one of the most effective treatments for major depressive disorder (MDD). There is recently increasing attention to evaluating ECT's effect on resting-state functional magnetic resonance imaging or rs-fMRI. This study aims to compare rs-fMRI of MDD patients with healthy participants (HC), investigate whether dynamic functional connectivity (dFC) estimated from rs-fMRI predicts the ECT outcome, and explore the effect of ECT on brain network states.
Methods: Resting-state fMRI data were collected from 119 MDD patients (76 females), and 61 HC subjects (34 females) with age mean of 53.46 (N=180) years old. The pre-ECT and post-ECT Hamilton Depression Rating Scale (HDRS) of MDD patients were 25.59±6.14 and 11.48±9.07, respectively. Twenty-four independent components from default mode (DMN) and cognitive control network (CCN) were extracted using group-independent component analysis form pre-ECT and post-ECT rs-fMRI. Then, the sliding window approach was used to estimate the pre-and post-ECT dFC of each subject. Next, k-means clustering was used to put dFC of all subjects in three distinct states. The optimum number of states estimated based on elbow criteria We calculated the amount of time each subject spends in each state, called occupancy rate or OCR. Next, we compared OCR values between HC and MDD participants in pre-ECT. We also calculated the partial correlation between pre-ECT OCRs and HDRS change while controlling for age, gender, and site. Finally, we evaluated the effectiveness of ECT by comparing post-ECT OCR of MDD and HC participants.
Results: The main findings include: 1) HCs had significantly higher OCR values than the MDDs in state 2, where connectivity between CCN and DMN was relatively higher than other states (corrected p=0.03), 2) Pre-ECT OCR of state 1, with more negative connectivity between CCN and DMN components, predicted the HDRS changes (R=0.22, corrected p=0.03). This means that those MDD patients who spend less time in this state showed more HDRS change with ECT, and 3) The post-ECT OCR analysis suggested that ECT increased the amount of time that MDD patients spend in state 2 (corrected p=0.03).

Conclusion: Our finding suggests that the dFC features, estimated from CCN and DMN, as a predictive biomarker of the ECT outcome of MDD patients. Also, this study presented an underlying mechanism associated with the ECT effect in MDD patients.

Abstract 19

The link between brain functional network connectivity and genetic risk of Alzheimer's disease

Mohammad. S. E. Sendi[1,2,3], David H Salat[4,5], Vince D Calhoun*[1,2,3,7]

[1]Wallace H. Coulter Department of Biomedical Engineering at Emory University and Georgia Tech, Atlanta, GA, USA
[2]School of Electrical and Computer Engineering (ECE) at the Georgia Institute of Technology, Atlanta, GA, USA
[3]Tri-Institutional Georgia State University/Georgia Institute of Technology/Emory University Center for Translational Research in Neuroimaging and Data Science, Atlanta, GA, USA, [4]Neuroimaging Research for Veterans Center, VA Boston Healthcare System, Boston, MA, USA
[5]Massachusetts General Hospital, Charlestown, MA, USA
[6]Harvard Medical School, Boston, MA, USA
[7]Georgia State University, Atlanta, GA, USA

*Corresponding Author

Background: Apolipoprotein E polymorphic alleles are genetic factors linked to Alzheimer's disease (AD). Individuals carrying the ε4 allele have the highest risk of AD compared with those carrying ε3 and ε2 allele, whereas ε2 allele has the lowest risk. Although previous studies explored the link between the genetic risk of AD and static functional network connectivity (sFNC), limited studies have evaluated the association between dynamic FNC (dFNC) and AD risk. Here, we explore how the dFNC differs between individuals with genetic risk for AD.

Methods: Resting-state fMRI (duration: 6 min) data of 991 healthy (clinical dementia rating=0) brains (404 females) and their demographic information from the longitudinal Open Access Series of Imaging Studies (OASIS)-3 cohort were used. The participants' age at scanning time was ranging from 42.46 to 95.39, with a mean of 69.81. We put the data into three groups including group1 (N=135, 63 females) including subjects with ε2 allele (i.e., ε2/ ε2, ε2/ ε3, and ε3/ε2), group2 (N=558, 219 females) including subjects with only ε3 allele (i.e., ε3/ ε3), and group3 (N=298, 122 females) including subjects with ε4 allele (i.e., ε3/ ε4, ε4/ ε3, and ε4/ε4). Age and gender were not significantly different across groups. Group independent component analysis was used to extract 53 data-driven components for the whole-brain. The sliding window and Pearson correlation were used to measure the dFNC among 53 components. A k-means algorithm was applied to the dFNC windows to partition them into three sets of separated states We calculated each subject's occupancy rate (OCR) in each state. A two-sample t-test was used to compare the OCR of groups in each state.

Results: Subject with a lower AD risk spend more time in state1 with more positive connectivity within cognitive control network (CCN) and between CCN and sensory network (corrected p<0.05). Interestingly, in this state, the difference of OCR among subjects with different AD risk was more significant in females, while males did not show any significant difference in their OCR across three groups. Females with higher AD risk had more OCR in state 3 with relatively lower within CCN connectivity.

Conclusions:

Results support the use of dFNC features as a potential biomarker of AD genetic risk for females. We also showed the role of CCN connectivity associated with the AD risk. This study is the first study that exploring links between dFNC and AD genetic risk to the best of our knowledge.

Abstract 20

A graph neural network model to estimate cell-wise metabolic flux using single cell RNA-seq data

Wennan Chang[1,2], Norah Alghamdi[1], Pengtao Dang[1,2], Xiaoyu Lu[1], Changlin Wan[1,2], Silpa Gampala[3], Yong Zang[1,5], Melissa Fishel[3]*, Sha Cao[1,5]*, Chi Zhang[1,2]*

[1]Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics
[2]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN 46202, USA
[3]Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
[5]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

*Corresponding Authors

The metabolic heterogeneity, and metabolic interplay among cells in the tissue microenvironment have been known as significant contributors to disease treatment resistance. However, with the lack of a mature high-throughput single cell metabolomics technology, we are yet to establish systematic understanding of the intra-tissue metabolic heterogeneity and cooperative mechanisms. To mitigate this knowledge gap, we developed a novel computational method, namely scFEA (single cell Flux Estimation Analysis), to infer cell-wise fluxome from single cell RNA-sequencing (scRNA-seq) data. scFEA is empowered by a systematically reconstructed human metabolic map as a factor graph, a novel probabilistic model to leverage the flux balance constraints on scRNA-seq data, and a novel graph neural network-based optimization solver. The intricate information cascade from transcriptome to metabolome was captured using multi-layer neural networks to capitulate the non-linear dependency between enzymatic gene expressions and reaction rates. We experimentally validated scFEA by generating an scRNA-seq dataset with matched metabolomics data on cells of perturbed oxygen and genetic conditions. Application of scFEA on this dataset demonstrated the consistency between predicted flux and the observed variation of metabolite abundance in the matched metabolomics data. We also applied scFEA on five publicly available scRNA-seq and spatial transcriptomics datasets and identified context and cell group specific metabolic variations. The cell-wise fluxome predicted by scFEA empowers a series of downstream analysis including identification of metabolic modules or cell groups that share common metabolic variations, sensitivity evaluation of enzymes with regards to their impact on the whole metabolic flux, and inference of cell-tissue and cell-cell metabolic communications. The key contributions of this work include: (1) scFEA is the first computational method to bridge the knowledge gap of high-throughout single cell metabolomic profiling by estimating cell-wise metabolic fluxome from scRNA-seq data, (2) scFEA is empowered by a novel graph neural network architecture and solver to model mass carrying flux over the complex whole metabolic network, (3) scFEA and its downstream analysis enable a comprehensive inference and characterization of metabolic stress, metabolic shifts and other biochemical contexts in tissue microenvironment.

Abstract 21

IPDB: Integrated Pregnancy Database with clinical and omics data

Parth G Kothiya[1], Huanmei Wu[1,2], David M. Haas[3], Shelley D. Dowden[3], Bobbie N Ray[3], Sara K Quinney*[1,3]

[1] Department of BioHealth Informatics, Indiana University, Indiana University–Purdue University Indianapolis, IN USA
[2] Department of Health Services Administration and Policy, Temple University College of Public Health, Philadelphia, PA USA
[3] Department of Obstetrics & Gynecology, Indiana University School of Medicine, Indianapolis, IN USA

*Corresponding Author

Introduction: Providing integrated electronic health records (EHR) and biospecimen data from pregnant women for bioinformatics analyses has the potentials to enhance knowledge and care. The heterogeneous and longitudinal nature of obstetric data makes integration challenging. To date, we are unaware of existing databases that combine EHR data with 'omics data longitudinally across pregnancy and postpartum. We have developed the Integrated Pregnancy Database (IPDB) to address these challenges for pregnancy data management and analysis.

Resources and methods: The IPDB source data include (i) Redcap data manually extracted from EHR of two local hospitals that comprise various clinical and demographic data (~600 clinical variables), including vital signs, past and current diagnoses, family and social history, medications, laboratory values, and maternal and neonatal outcomes; (ii) biobank samples and resulting omics data and other laboratory tests; and (iii) longitudinal data, such as smoking and drinking before and during pregnancy and clinical visits. Data cleaning, preprocessing, reformatting, normalization, and standardizing are first performed before data merging.
The IPDB was built with a multi-modal concept using a relational database design to link the heterogeneous information. The core database tables are pregnancy, mother, clinical_visit, biospecimen, and neonatal_outcomes. Special entity and relationship tables are carefully designed for complicated situations, such as one pregnancy with multiple babies, multiple pregnancies, abnormal birth outcomes, and connecting EHR data with biosample analysis. The other design principle is the extensibility of the database, which can easily incorporate additional clinical variables, extra lab tests, and new analytical results. Yet database design considerations include scalability to allow new datasets, intensive query processing, data access controls, and system recovery.
Results: We have developed a prototype of the IPDB using a MySQL database that integrates longitudinal clinical information, biosample inventories, and omics data analyses (metabolomics, genomics, etc.). Users can access and manipulate data through a graphical user interface (GUI) implemented by R-shiny. Users can conveniently query information over multiple pregnancies and evaluate various outcomes. Moreover, it empowers users to obtain postpartum information and laboratory values. A visualization dashboard has been developed with descriptive statistics and illustrations over selected data on maternal age, gestational age, race, health factors, and outcomes.
Summary: The cross-modality feature of IPDB integrates multiple data streams, from structured EHR data to genomics and metabolomics, in a format facilitating future bioinformatics analyses. It can significantly improve the performance of phenotyping and prediction algorithms, enabling knowledge discovery at the patient and population level.

Abstract 22

Language of the transcribed enhancers using NLP algorithm

Karla Paniagua[1], Mario Flores*[2]

[1,2]Department of Electrical Engineering, University of Texas at San Antonio, San Antonio, TX, USA.

*Corresponding Author

A proportion smaller than 2% of the mammalian genome codes for proteins, hence, the search for noncoding functional DNA is a subject of interest in biological science. In this

investigation functional categories of transcribed enhancers will be investigated genome wide. Recent studies demonstrated, using next generation sequencing, that transcribed enhancers are distributed along the human genome, interacting with different molecules and compounds. The main goal is to understand the language of transcribed enhancers and their task within cellular functions. With the support of NLP, an algorithm will be trained using these categories of transcribed enhancers to identify a network of proteins, giving us a better insight into the function of these regulatory proteins.

Keywords: DNA, Enhancer, NLP, proteins, transcription

Abstract 23

Detection of DNA replication origins via peak calling of DNA modifications

Gogoate Lemea[1], Qian Liu[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, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

*Corresponding Author

Background: Replipore sequencing enables the detection of modified nucleotides incorporated during DNA replication. However, genome-scale identification of DNA replication origins from replipore sequencing is still a challenging problem.
Methods: In this study, we propose a "modification trend" approach to detect DNA replication origins from Nanopore sequencing. After a fraction of Thymidine analogs (a type of modification) are introduced in DNA replications and sequenced via Nanopore sequencing, a peak calling method is designed to capture the trends of how Thymidine analogs changes cross a chromosome. To reduce random errors, smoothing window is used to generate median incorporation percentage of Thymidine analogs for a small region, and peaks are called via Pearson correlations for local regions. Each peak indicates a local maximum incorporation of Thymidine analogs, suggesting a potential location of DNA replication origins.
Results: We tested this peak calling process on 3 DNA modification datasets of *Saccharomyces cerevisiae*, and compared called peaks of Thymidine analogs with wet-lab determined replication origins from OriDB. We found that on each dataset, the majority (~90%) of called peaks are adjacent to confirmed and likely origins from OriDB, although less than half of replication origins are recalled. We then combined peak calling results from different datasets to improve recall rates.
Conclusions: We proposed a peak calling method to identify DNA replication origins at a genomic scale from predicted DNA modifications via Nanopore sequencing. We expect that the method may be applied on human samples to facilitate disease studies in the future.

An interactive shiny web application for exploring a deep learning-predicted cancer dependency map of tumors

Gabriela Rubannelsonkumar[1,2], Yu-Chiao Chin[1], Yidong Chen*[1,3]

1 Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA
2 Bachelor of Science in Bioinformatics Program, Department of Biological Sciences, St. Mary's University, San Antonio, TX 78228, USA
3 Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA

*Corresponding Author

The Cancer Dependency Map (DepMap) project utilized genome-wide loss-of-function screens to quantify the degree to which a gene is essential for cancer cells' growth and survival. We recently published a deep learning model based on a unique transfer learning design, namely DeepDEP, that extended cell line-derived DepMap data to tumor's context by predicting tumors' gene dependencies using genomic profiles. Using the DeepDEP model, we constructed a pan-cancer dependency map of 8,238 tumors of The Cancer Genome Atlas (TCGA). Yet, an interactive platform to explore the predicted cancer dependency map of tumors remains to be developed. Here we present an R shiny server that allows easy querying of the cancer dependency data and their association with patients' clinical parameters and survival outcome without the prerequisite of computational expertise. The web application consists of four main modules: (1) query of a tumor or a gene of interest, (2) query of a cancer type of interest, (3) association analysis with baseline clinical variables, and (4) survival analysis of cancer dependencies across individual cancers. Module 1 allows users to search a specific tumor with the TCGA tumor barcode and rank all its gene dependencies from strongest to weakest. Users can also generate a pan-cancer dependency plot of a gene of interest. Module 2 allows the user to choose a cancer type and then sort dependencies based on the significance of difference between the cancer of interest and others. The results are visualized by a heat map. Module 3 tests the associations between the dependency score of a gene and baseline clinical variables (e.g., age, gender, race, ethnicity, TNM stage, pathologic stage, cancer-specific markers, etc.). Users can easily choose a clinical variable of interest and an interactive plot. Module 4 generates Kaplan-Meier curves of each gene dependency in individual cancer lineages to understand the potential of cancer dependencies as prognostic markers for overall survival. Users can set the cut-offs to define strong and weak dependency for each gene, with a default of a median cut-off. In summary, the present shiny web application makes the DeepDEP-predicted tumor dependency data readily available for biomedical researchers,

enabling comprehensive analysis of the dependency data with clinical data that may lead to clues of potential therapeutic and prognostic targets.

Abstract 25

Therapeutic Re-Positioning of Amiloride: From Anti- hypertension to Anti-Cancer

Aleshia Seaton-Terry [1], Venkataswarup Tiriveedhi*[2]

[1]Department of Biological Science, Tennessee State University, Nashville, TN, USA
[2]Department of Biological Science, Tennessee State University, Nashville, TN, USA

*Corresponding Authors

High salt concentration is recognized as being an initiator for the pro- inflammatory cascade in cancer. Previous studies demonstrate high salt concentration plays a direct role in cancer progression by inducing angiogenesis and immune dysfunction. Our laboratory has shown that breast cancer cells exposed to high salt concentration release inflammatory cytokines (TNFα and IL1β) by epithelial sodium channel (ENaC) mediated signaling. Moreover, our lab has revealed a correlation between high salt concentration and up-regulation of ENaC in breast cancer. Indeed, ENaC is recognized as a regulator of breast cancer proliferation and metastasis. Findings regarding ENaC have suggested ENaC as a potential drug target in the therapeutic treatment of breast cancer. However, previous research has failed to address potential drug inhibitors of ENaC that prevent transcription factors responsible for up-regulating pro-inflammatorysignaling cascade in breast cancer. In our current study, we attempt to understand Amiloride, an ENaC inhibitor, interaction with ENaC as a possible inhibitor of the inflammatory response in breast cancer cells. Our study focuses on Beta2- Adrenergic Receptor (ADRB2) a known cell membrane G Protein Coupled Receptors (GPCR) which are responsible for inflammatory signals that results in chronic stressed induced tumor progression and metastasis. Studies have revealed Beta2-Adrenergic Receptor (ADRB2) possess oncogenic activity and shows over-expression in numerous cancers including breast cancers. It has not yet been established whether Amiloride has an inhibitory effect on ENaC downstream signaling in breast cancer. In this study, using molecular modeling approaches, we focused on in silico identification of binding domains between Amiloride, ENaC-ADRB2 that lead to activation and downstream signaling of transcription factors that result in inflammatory cytokine production. Furthermore, we examined the potential anti-tumor effect of Amiloride through inhibition of ENaC- ADRB2 interaction. Our studies will provide evidence for a novel re-positioning of Amiloride, an anti-hypertensive drug, as a possible promising anticancer therapy.

Abstract 26

Druggability Predictions on a Proposed Common Allosteric Binding Site of G-Protein Coupled Receptors

Faisal Malik[1], Zhijun Li*[1]

[1]Department of Chemistry and Biochemistry, University of the Sciences, Philadelphia, PA, USA.

*Corresponding Author

Mediating a large portion of responses to extracellular stimuli, G-protein coupled receptors (GPCRs) are the largest group of integral membrane proteins responsible for cross-membrane signaling transduction. These receptors are allosteric in nature and have been identified to bind with a series of individual positive and negative allosteric modulators. The hypothesis of this study centralizes on the proposition that there is a common allosteric binding site located near the G-protein binding site in all GPCRs, regardless of the receptor's classification. To test this, Schrodinger's SiteMap module was used to the to detect the allosteric binding site in a series of experimental structures documented in the Zhang Lab database. Given that the allosteric site was present in all the experimental structures, PockDrug, a pocket druggability prediction webserver, was employed to determine each pocket's druggability. From these experimental steps, it was determined that this allosteric site is present in all GPCRs and the majority of sites identified by SiteMap were deemed druggable. Current steps in the procedure aim to measure the conservation levels of this allosteric site in each of the receptors by applying a conservation analysis. This will provide necessary evidence that the allosteric site is a selective site which will be useful in eventual drug design applications. Allosteric modulators play an important and specific role in pharmacology, as they allow for a greater selectivity and efficacy compared to orthosteric drugs. Applying these approaches will expand on the use of allosteric modulation in the treatment of various diseases associated with GPCRs. GPCRs are major drug targets, and allosteric drugs provide a level of improvement over traditional drugs that work at the primary endogenous binding site, specifically for those GPCRs that cannot be targeted traditionally.

Abstract 27

Delineating cell state heterogeneity in bladder cancer

Antara Biswas[1], Sivasomasundari Arunarasu[2], Subhajyoti De*[1]

[1]Rutgers Cancer Institute, Rutgers the State University of New Jersey, New Brunswick, NJ 08901, USA
[2]Emory University Atlanta, GA 30322, USA

*Corresponding Author

Heterogeneity and evolvability are hallmarks of all cancers. Despite sharing their evolutionary history from a single somatic cancer initiating cell, tumor cells tend to have genetic and nongenetic variations among themselves. Although some of these variations are inconsequential, others tend to contribute to cell state transition and phenotypic heterogeneity, providing a substrate for somatic evolution. Some aspects of intra-tumor heterogeneity might be reversible i.e. tumor cell phenotypes can dynamically change under the influence of genetic mutations, epigenetic modifications, and microenvironmental contexts - which in turn, can confer resistance to treatment, promote metastasis, and enhance evolvability in cancer. We develop a computational and genomic framework to infer the rate of dynamic cell state transition in cancer and also in vitro model systems based on imaging and single cell genomic data. Using this framework, we determine the rate and drivers of dynamic cell state transition in a cohort of human bladder tumors. We then model similar cell state transition in a bladder cancer cell line model under controlled laboratory condition. We further determine the change in tumor cell state dynamics under stress associated with nutrient starvation, akin to that expected in tumor microenvironment, and also during radiation treatment. We propose that dynamic heterogeneity shapes tumor evolution and can modulate efficacy of anti-cancer treatment.

Abstract 28

Construction of Homogeneity-Inspired Neural Network for Nonlinear Ordinary Differential Equations

Kelsey Mitchel[1], Shiqi Nan[1], Chunjiang Qian*[1]

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

*Corresponding Authors

The existing Neural Ordinary Differential Equation (Neural ODE) network was created based on the connection between the deep learning technique of neural networks and the modeling of dynamic systems. Neural ODEs are a continuous-depth model based on differential equations solvers, and are used to model complex dynamic systems, such as continuous time systems. This type of neural network is capable of modeling biomedical systems in order to predict medical outcomes that evolve over time, such as bone fragmentation. However, the Neural ODE's current structure is unable to accurately model some complex nonlinear dynamic systems. We created a Neural ODE structure that will better model specific continuous-time dynamic systems. Our initial findings have shown that using activation functions within the Neural ODE that have the same degree and a similar structure to the system being modeled results in better performance when compared to other activation functions.

Predicting COVID-19 Infection Severity Using CNN on Single Cell RNA-seq Data

Leonardo Falcon[1], Wenjian Huang[1], Karla Paniagua[1], Ricardo Ramirez[2], Mario Flories[1], Yu-Fang Jin*[1]

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

*Corresponding Author

Coronavirus Disease 19 (COVID-19) infection has demonstrated a highly variable severity among the infected patients. Some patients experience mild to no symptoms while others come down with severe illness, leading to one of the research problems to be investigated for COVID-19 treatment. The goal of this study is to uncover underlying regulatory mechanisms for different severities of COVID-19 infection using single-cell RNA sequencing (scRNA-Seq) data and deep learning algorithms. In particular, scRNA-Seq data collected from severe and mild cases were compared to uninfected cases using convolutional neural network (CNN) approaches. Raw datasets of previously published scRNA-Seq of bronchoalveolar lavage fluid were aggregated, normalized, and quantified. The raw dataset includes 23,071 cells from uninfected healthy control, 12,598 cells from patients with mild symptoms, and 47,461 cells from patients with severe symptoms, leading to a total of 83,130 cells and expressions of 32,738 genes. The filtered dataset includes 54,214 cells (17,480 cells from uninfected healthy controls, 2502 cells from patients with mild symptoms, and 34,232 cells from patients with severe symptoms), and expressions of 14331 genes. To examine the properties of the filtered dataset, expression levels of genes were clustered and presented using the Uniform Manifold Approximation and Projection (UMAP) method.Two different CNN models were established with different embeddings. The embedding approach redefined each gene as a pixel and the expression level of the gene was assigned to that pixel. Through the embedding process, the 1-D gene expression of a cell was represented as a 2-D image of the cell for input to CNN models. The first CNN model was established by randomly positioning a cell's genes into an image while the second CNN model was established by positioning genes into an image by their chromosomal order. The CNN models had one convolutional layer, one hidden layer with 1,024 nodes which were fully connected to the output layer. The output layer included three nodes representing 3 different severities of COVID-19 infection: uninfected, mild, and severe. Cells from each case of severity were partitioned 80% for training, 10% for testing, and 10% for model validation. The CNN models predicted cells from patients with uninfected, mild, and severe symptoms with an accuracy of 98-99%.

The learned CNN features in the hidden layer were extracted to identify the leading genes and functional modules for different severities of COVID-19 infection.

Abstract 30

Data-Driven Control Strategies for Largescale Multi-Agent Systems

Carl Arthur Jones[1], Claire Walton*[1.2]

[1]Department of Mathematics, University of Texas at San Antonio, San Antonio, TX, USA [2]Department of Electrical & Computer Engineering, University of Texas at San Antonio, San Antonio, TX, USA

*Corresponding Author

Molecular interactions are important in biological processes and have attracted extensive research effort to study the behavior of molecules. Multi-agent modeling is one of the promising research directions to simulate and study the behavior of molecular interactions, where each molecule can be defined as an agent. Representing thousands of molecules in a biological process imposes significant computational and theoretical challenges. Control of these largescale systems—also called `swarms' in multi-agent modeling—is an additional challenge. Large-scale cooperative control algorithms for guiding swarm behavior have been of particular interest in recent years due to widespread applications, including molecular dynamics, bioinformatics, medicine, particle swarm optimization, and intelligent computing research. Currently, control and regulation of multi-agent systems with even greater than one hundred agents is very difficult. Optimization may assist with this difficulty. For example, recent work has shown that optimizing input parameters for particle swarm optimization improves the speed of these algorithms and the efficiency at which such an algorithm can train a neural network to learn a predictive model of blood-brain barrier permeation. Many control inputs and parameters remain to be targeted for optimization in these models. For instance, particle swarm optimization performance is largely dependent on swarm size (number of agents), and this number is often a matter of ad hoc decision making per intended application. We investigate the optimization of swarm cooperative control through the study of two models in the literature, both of which have the capability of modeling aspects of intermolecular dynamics: the Reynold's boid model, and the virtual body artificial potential model. We perform optimization by identifying parameters which lead to optimized swarm behavior, measured by key behavior metrics. We then utilize nonlinear programming techniques for large-scale systems and differential equation approximation methods to optimize parameter values. We consider optimized parameters for different performance metrics and also analyze the impact of swarm size on optimal parameter values.

NOTES

About IAIBM

IAIBM

The International Association for Intelligent Biology and Medicine (IAIBM, http://iaibm.org/) is a non-profit organization. It was formed on January 19, 2018. Its mission is to promote the intelligent biology and medical science, including bioinformatics, systems biology, and intelligent computing, to a diverse background of scientists, through member discussion, network communication, collaborations, and education.

Special Acknowledgments

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

Khandakar Tanvir Ahmed, University of Central Florida, USA

Jingxuan Bao, University of Pennsylvania, USA

Qibing Jiang, University of Central Florida, USA

Jim Havrilla, Children's Hospital of Philadelphia, USA

Yu Hu, Children's Hospital of Philadelphia, USA

Chong Li, Temple University, USA

Andi Liu, The University of Texas Health Science Center at Houston, USA

Mohammad Erfan Molaei, Temple University, USA

Jiao Sun, University of Central Florida, USA

Maria Vazquez Velazquez, The University of Texas Health Science Center at Houston, USA

Zixuan Wen, University of Pennsylvania, USA

MANY THANKS TO OUR SPONSORS!